Automated Marking of Exam Papers

Using Semantic Parsing

Jamie Frost



Contents

| 1. | Introduction | , | | | |
|-----|---------------------------------------|---|--|--|--|
| | Motivation | *************************************** | | | |
| | CAndC Parser & Boxer | | | | |
| | The Data | | | | |
| | Weka | | | | |
| 2. | System Overview | | | | |
| 3. | Parsing | | | | |
| 4. | Representing Semantic Atomic Facts | | | | |
| | Relevance to Exam Marking | | | | |
| 5. | Data Preparation and Training | | | | |
| 6. | Spelling Correction | | | | |
| | Dictionary Comparison | | | | |
| | Corpus Comparison | | | | |
| | Other Considerations | | | | |
| | Mapping the Corrections | | | | |
| | A Demonstration | | | | |
| 7. | Cumulative Classification | | | | |
| | Segmentation | | | | |
| | Method 1: Standard K-means Clustering | | | | |
| | Method 2: Iterative Clique Merging | | | | |
| | Method 3: Manual specification | | | | |
| | Training | | | | |
| | Classification | | | | |
| | A worked example | | | | |
| | Pre-Testing Analysis | | | | |
| 8. | Graphical User Interface | | | | |
| 9. | Evaluation | | | | |
| | Results | | | | |
| | Cumulative Classifier | 28 | | | |
| | Spelling Correction | | | | |
| 10. | Conclusion | 30 | | | |
| 11. | Implementation notes | 31 | | | |
| | a) Parsing | 31 | | | |
| | b) Implementing $arphi$ | 31 | | | |
| | c) Dictionary Methods | 33 | | | |
| | d) Cumulative Classification | 33 | | | |
| 12. | Bibliography | 34 | | | |
| 13. | Acknowledgements | 34 | | | |

1. Introduction

Motivation

'Automated Essay Scoring' has been a large area of research since the 1960s. In such a process, a variety of 'features' are extracted from essays, such as word and sentence length and the structure of sentences, before the data is collaborated to provide a final classification [SHERMIS 03].

'Automated Exam Scoring' is a more objective mode of classification, in which answers are analysed for the presence of concrete facts or statements instead of using any continuous measure. This has been the subject of much research in the Computational Linguistics department at Oxford University, using an online study as the source of data, in which students completed a GCSE Biology paper [Pulman 05]. The techniques employed are varied, but falls into two main categories. One simulating a human style marker defines the marking scheme via patterns inputted by an administrator, allowing for as many variants of an answer as possible. The clear disadvantage of this method is the hours of work required to painstakingly write these patterns, but this method yields high accuracy. Average accuracy in excess of 95% was obtained.

The latter method adopts a machine learning approach using a set of pre-marked answers for the training process. [PULMAN 06] experimented with a system in which the answers are treated as a 'bag of words' with no semantic structure incorporated. A technique known as 'k nearest neighbour (KNN)' was used, which initially removes stop words from the answer sets (words of common usage such as 'and') and assigns each keyword a TDF-IDF measure ("Term Frequency" and "Document Frequency", concepts in the Information Retrieval domain combined to indicate the 'significance' of a term). Each answer in the training set is then converted to an incidence vector form given its terms (weighting each element with the corresponding term's TDF-IDF measure), as well as the answer to classify. Using a distance measure, the closest k training answers are obtained, and the modal score of these used for classification.

This naive method is subject to a number of problems, as highlighted by Professor Stephen Pulman in this BBC News article:

"Prafessor Srihari asked human examiners to grade 300 answer booklets. Half af the graded scripts were then fed into the computer to "teach" it the grading process. The software identified key words and phroses that were repeatedly associated with high grades. If few of these features are present in an exam script, it generally receives a law grade...

Prafessor Stephen Pulman at the University of Oxford has identified another potential pitfall in Prafessor Srihari's approach. "You can't just look for keywards, because the student might have the right keywords in the wrong configuration, or they might use keywords equivalents," Professor Pulman says."

Thus if the answer requirement is a statement such as 'the cat chased the mouse', then an answer of 'the mouse chased the cat' would be accepted despite the clear semantic inequality, due to the identical set of words.

This project aims to extend this method by incorporating the semantic structure of sentences, so that for the above example 'the mouse chased the cat' would be marked as incorrect, whereas 'the mouse was chased by the cat' would be marked as correct.

CAndC Parser & Boxer

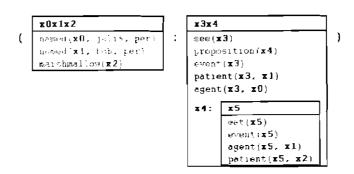
The CAndC (Clark and Curran) parser uses statistical methods and 'supertagging' to convert English sentences into a tree representing their structure, as detailed in [CLARK 07]. It was developed by Stephen Clark (a lecturer at Oxford University and member of the Oxford Computational Linguistics Group) and James Curran. The output of the parser is a CCG (Combinatory Categorial Grammar) file, representing this sentence structure.

Alone, this representation is insufficient for machine learning use, given that the semantic interpretation of the sentences is our concern. We therefore use a tool called *Boxer*, which uses Prolog to convert the CCG into a form called DRS (Discourse Representation Structure). This is compatible with first-order logic, and thus can be used to make reasoned logical deductions (with its application extending to other systems such as Question Answering).

Consider the sentence "Julia saw Bob eat the marshmallow." Its CCG form is:

| Julia NNF N | \$aW VBD ((S[dil] (IP) (S[5 NP)) NP | Bob | eat VB (S(b) NP/NP | the DT NP(nb) N | mar shmallow NN N | | | |
|-------------------|---|-----|--------------------------|-----------------------|-------------------|--|--|--|
| | (3[3:1] SP), (3[3] NP | | 2[P] NP | | | | | |
| | Hocker | | | | | | | |
| | <u></u> | 5 | [6d] | | | | | |
| | | | 5(8:13 | | <u> </u> | | | |

The graphical form of its corresponding DRS is:



Verbs and nouns are represented as objects (here x0 to x5) with predicates defining their properties and relations connecting them.

The Data

All data is from students' answers to GCSE Biology papers (both foundation and higher tiers) from 2001 and 2003. The anonymised data is made available under a confidentiality agreement courtesy of the OCR Exam Board. Maximum marks for answers vary between 1 and 4 marks.

Weka

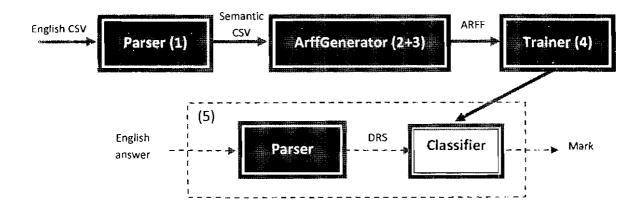
Weka is a suite of machine learning tools for Java, developed by the University of Waikato in New Zealand. It provides a variety of classifiers (including Naive Bayesian, Neural Networks and Decision Trees), clustering algorithms and other utilities. Downloads and documentation can be found at http://www.cs.waikato.ac.nz/~ml/weka/.

2. System Overview

In the scope of this project, we investigate 3 different methods of exam mark classification: a naive keyword approach in which the answers are treated as a 'bag of words' as described above (but using a more standard classifier than the 'k nearest neighbour' technique), an approach which incorporates the sentence semantics, and a potential optimisation of the latter which uses a new type of classifier. All 3 methods use a 'supervised learning' approach, that is, we train a classifier with pre-marked data, and use this to mark subsequent answers.

Such a classification problem requires a number of stages:

- Semantic Conversion: This takes the answers in English (a CSV file containing the answers and marks), and converts them to a semantic representation, using the CAndC parser to identify the sentence structure. This is then fed this into Boxer, producing a semantic representation (i.e. a DRS). Clearly this step is skipped in the keyword approach. The conversion process is described in Section 3.
- 2. Featurisation: A 'feature' is a singular property encompassing some fact about an answer. A 'binary feature' takes the value 0, denoting the feature is not present, or 1 if it is. The aim of this stage is to provide some function which maps the answer (either in its raw form for the keyword approach, or the DRS for the semantic approach) to a set of such 'features' which are satisfied. For the keyword approach, the mapping simply produces a set of words the answer contains. For the semantic approach, the DRS is a non-linear structure and thus the mapping is more ambiguous. The latter is dealt with in Section 4.
- 3. Data Preparation: Such features from all answers must be collated to provide a suitable input for a classifier. This data is embodied as a file known as an 'ARFF'. See Section 5.
- 4. Training: A suitable machine learning classifier is chosen, and is 'trained' with this data.
- 5. Classification: Subsequent new answers are parsed (if necessary) and featurised as before. This is fed into the trained classifier to provide an estimated classification (i.e. a mark) for the answer.



3. Parsing

Our initial input data is a CSV file containing a list of English sentences, each with a classification. The aim of the parsing process is to produce a new CSV file, instead with DRS expressions representing each of the sentences. There are 2 methods of parsing.

One method uses the online parser found on the CAndC website.¹ However, a more practical solution would incorporate a server which can administer requests from clients and send the parsed expressions back. The CAndC tools include a *SOAP* server and client, which communicate with each other (potentially externally). The advantage of these is that the high overhead cost associated with starting the parser need only occur once in the server's lifetime, relative to once per parse request. Both of these operate on the same host, and we disregard their remote capabilities of these in order to produce a Java 'wrapper' which gives this functionality. The *SOAP* client should not be confused with the remote Java client, the latter of which generates the parse requests.

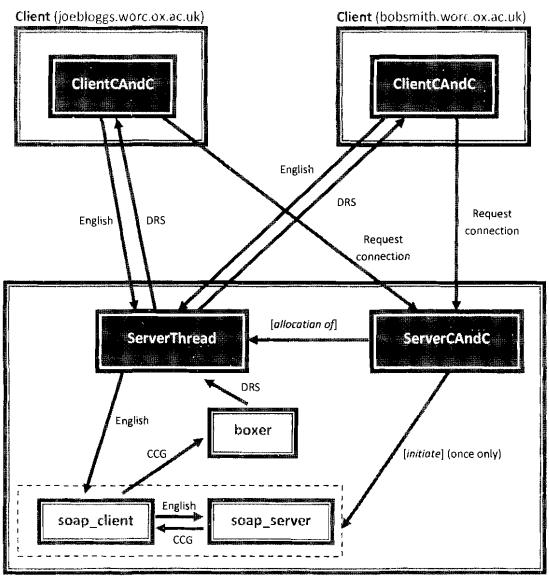
What is required is a server which forwards requests to the SOAP client (which in turn communicates with the SOAP server). It may need to manage the requests of multiple remote clients, and thus we create a separate manager for each connection, running concurrently. It performs the following:

- 1. The SOAP server is started.
- We indefinitely wait on a specified port for external clients to make a connection. When an incoming request is received, a socket and a manager representing the connection are created.
- 3. In this manager, we construct an input and output stream from the socket, and continually read from it until the connection closes. We read an English sentence from the stream, write it to a file (given that the parser reads from a file) and instruct the SOAP server to parse the sentence. Finally we use Boxer to convert this to a DRS expression, which is delivered to the output stream.

A client end of the communication is also required. It sends the sentence to the server and delivers the parsed and boxed DRS expression returned to the required resource. As before, we establish a socket using the host name and port of the external machine running the server, and construct an input and output stream to read and write from.

The structure of the entire parsing system can be summarised as follows:

¹ This can be found at http://svn.ask it.usyd.edu.au/trac/candc/wiki/Demo



Server (curlew.comlab.ox.ac.uk)

4. Representing Semantic Atomic Facts

As previously discussed, previous attempts at implementing a machine learning system to classify answers have relied on treating them as a 'bag of words', such that marking was based on common keywords. However, this has a major flaw. The ordering and form of words have no bearing, leading to false positives. For example "Jack collects stamps with Jill" and "Jill stamps on Jack as he collects Jill" are likely to have similar classifications, despite their very different meanings and different uses of the word 'stamps'. In other words, there is no capacity for semantic representation. In the latter sentence, although Jill is performing the 'stamping', it could equally be Jack.

The aim therefore is to derive a way of semantically representing sentences that is complete, sound and most importantly, linear. That is, we wish to deconstruct a sentence into a collection of atomic predicates (here described as 'semantic atomic facts') such that these properties are satisfied. This is an example of 'featurisation', that is, we identify a number of possible properties of the input data (in this case, the presence of each of the generated atoms) for the use of machine learning application at a later stage. To establish these concepts of soundness, completeness and linearity, take for example a sentence represented by α and the 'semantic atom facts' we wish to generate from it by a set $\{\beta_1, \ldots, \beta_n\}$:

- For soundness, $\forall i \cdot \alpha \vDash \beta_i$. In other words, all such facts are implied by the original sentence.
- For completeness, $(\land |\lor | \neg)$ $\beta_i \models \alpha$. That is, some disjunctive/conjunctive/negative combination of the atoms implies the original sentence. Allowing the disjunctive connective to yield α may at first seem odd, but the necessity becomes clear when considering sentences with the word 'or'; if for example "The kite is red or green" was split into "Red kite" and "Green kite".
- Linearity simply implies the lack of some recursive or complex data structure, instead keeping the atomic facts 'flat'. Here, atomicity implies the absence of the A, V and ¬ connectives.

We define a suitable function ϕ which produces every possible inference that can be made from a DRS expression α , that is, the β_i as defined above. Take for example the sentence "When jolly Carol laughs, there is an earthquake". We can make a number of inferences from the sentence, representing looser forms:

- 1. When Carol laughs, some event occurs.
- 2. When someone laughs, some event occurs.
- 3. When Carol performs some action, some event occurs.
- 4. When jolly Carol laughs, some event occurs.
- 5. Carol is involved in the sentence.
- 6. Carol is jolly.
- 7. An earthquake occurred.
- 8. When someone laughs, an earthquake occurs.

And so on. To formally define ϕ , we consider all the mappings for all possible DRS sub-expressions to sets of 'atoms'. The DRS consists of a number of objects, representing nouns or verbs, named via a specially typed predicate. The atom produced is simply the name of the predicate (note that the term after the colon indicates the type):

```
\varphi(p:noun(x)) = { p:String }
 \varphi(p:verb(x)) = { p:String }
```

For example, $\varphi(\text{cat}(x)) = \{ \text{"cat"} \}$. In addition, other unary predicates representing adjuncts can modify such an entity; this is an adjective in the case of a noun, and an adverb in the case of a verb. Its corresponding atoms consist of the name of the modifier and the modifier combined with the name of the entity. The nm(x) function simply retrieves the assigned 'name' of the entity x.

```
\varphi(p:adj(x)) = { p:String[nm(x)]}
```

For convenience, we define $\varphi(x)$ to be all atoms associated with the entity x. We similarly take φ of the whole DRS expression α as the union of sets produced by applying the function to all expressions contained within it. That is:

```
\varphi(x) = \varphi(p_1(x); ...; p_n(x)) = \varphi(p_1(x)) \cup ... \cup \varphi(p_n(x))
\varphi(\alpha) = \varphi(\alpha_1; ...; \alpha_n) = \varphi(\alpha_1(x)) \cup ... \cup \varphi(\alpha_n(x))
```

Binary predicates (i.e. relations) connect two entities, and typically represent a verb or a prepositional modifier. In mind of providing all inferences, the resulting atoms must incorporate every possible pair of atoms from the atom sets of the constituent entities:

```
\varphi(p(x_1, x_2)) = \{p:String[e_1, e_2] \mid e_1 \in \varphi(x_1) \land e_2 \in \varphi(x_2)\}
```

Larger DRS expressions can be combined together using a variety of conjunctives. These must be dealt with appropriately:

```
\begin{array}{lll} \phi(\alpha\vee\beta\,) & = & \phi(\alpha)\cup\phi\,(\beta) \\ \phi(\alpha\wedge\beta\,) & = & \phi(\alpha)\cup\phi\,(\beta) \\ \phi(\alpha=/is\,\beta) & = & \{\mbox{$"e_1=e_2"$, $"e_2=e_1"$ | $e_1\in\phi(\alpha)\wedge e_2\in\phi(\beta)$ }\} \\ \phi(\alpha\to\beta) & = & \{\mbox{$"e_1\to e_2"$ | $e_1\in\phi(\alpha)\wedge e_2\in\phi(\beta)$ }\} \\ \phi(-\alpha) & = & \phi(\alpha) \end{array}
```

Such a function may initially seem disastrously hideous to a logician: negation is ignored, and conjunction and disjunction are treated in a uniform way. The motivation is to have as high coverage as possible of all inferences made by the sentence, with the only requirement being that the original sentence can be reconstructed from some conjunctive/disjunctive/negative combination of a subset of $\varphi(\alpha)$. Note that the equality relation is symmetric, therefore we must supply both orderings in the string representation. We do not treat $\alpha \to \beta$ as $\neg \alpha \lor \beta$ given we are only interested if the inference was made, not if it was true vacuously when α is false.

To demonstrate this definition, take the sentence "The big dog jumped over the lazy man." This could be represented in DRS by:

```
dog(x_1); big(x_1); man(x_2); lazy(x_2); jump(x_3); agent(x_3,x_1); over(x_3,x_2)
```

In this example, the jumping is being executed by the dog (i.e. the dog is the agent of the jumping). Using the laws above, the following atoms are produced:

dog
big
big[dog]
man
lazy
lazy[man]
jump

agent[jump,big[dog]]
agent[jump,dog]
agent[jump,big]
over[jump,lazy[man]]
over[jump,lazy]
over[jump,man]

Relevance to Exam Marking

The above system is based on assumptions relating to a generic marking system. We assume that some number of marks are awarded for the conjunctive/disjunctive combination of these atomic facts. For example, a mark may be awarded for mentioning that the man is lazy ("lazy[man]") and another for identifying a jump occurred over him ("over[jump,man]").

There may initially appear to be a problem with answers either being too specific or not specific enough. Say for example that "over[jump,man]" was required for a mark. If the user was to answer with "jumped over the large man", we obtain "over[jump,large[man]]" and by deconstruction, "over[jump,man]" also. Is a mark awarded for "jumped over the *large* man" if the answer required was "jumped over the man"? This is somewhat of an ambiguous question. In some cases, the extra detail may be unnecessary but not detrimental to the answer; we may consider that our example merits the mark. However, there are some cases where clearly extra detail changes the answer. Take for example the answer "The act of putting magnesium in water causes a reaction." Then "The act of putting magnesium in water causes a *fatal* reaction." is clearly wrong, given that it grossly misrepresents the magnitude of the reaction.

Fortunately, a suitable machine learning method can usually cater for this. Let an answer be α (e.g. red dog) and its looser implication β (e.g. dog) such that $\alpha \to \beta$ (i.e. a red dog is a dog):

- For an actual answer α where β is not specific enough (e.g. where *red dog* would be awarded a mark but not dog): $\alpha \land \neg \beta$ holds.
- For an actual answer β where α is an acceptable extension (e.g. if dog was accepted, then $red\ dog$ would also be accepted): $\alpha \vee \beta$ holds.
- For an actual answer β where α is not a valid extension (such as with the chemistry example above): $\neg \alpha \land \beta$ holds.

These are all simple Boolean expressions which can be 'learned' during the training process. Classifiers' ability to learn any such Boolean expression justifies our function ϕ used for linearisation.

There is however one particular flaw with the definition of ϕ , in the way that negation is handled. It is difficult to distinguish between "The dog was red" and "The dog was not red", since the negation is ignored. The underlying problem is that the binary value in the vector representation of the answer represents the presence or absence of a clause, not the Boolean interpretation of 'true' or 'false'. We could propagate the ~ through subclauses (resulting in sets like { dog, red, red[dog] } and { dog, ~red, ~red[dog] }), or use 3-valued logic (i.e. 0 for the absence of a feature, 1 if present but negative, and 2 if present and positive), but such attempts generally lowered the accuracy of the system (by up to 10% for some data sets). One might suggest that this is due to students generally avoiding specifying a correct answer before negating it. For example, the question "By which process to plants produce oxygen and food?" is unlikely to be answered with "Not by photosynthesis."

5. Data Preparation and Training

Parsing and atomising from the previous stage will produce sets of atoms for each of the answers. In order to produce an input file suitable for a classifier, these must be collated.

In Weka, a training file for a classifier is known as an ARFF (Attribute-Relation File Format) file. The file has two segments: the first is an ordered list of attributes (each with a data range defined), and the second a data table representing the instances (i.e. answers) as vectors, where each element represents a value for the corresponding attribute (the last value being the classification for the instance).

Producing the set of attributes is trivial; we simply take the union of all the sets of atoms. To produce the data table, for each instance in the training set we use the value '1' for an element if the attribute appears in its set of atoms, and '0' otherwise. This is best illustrated by an example; consider these 3 sets of atoms for the 3 answers in the training set:

```
A_1 = \{ \text{ cat, dog, mouse } \}, A_2 = \{ \text{ dog, horse, rabbit } \}, A_3 = \{ \text{ horse, monkey } \}
```

Suppose that A_1 received 1 mark, A_2 received 0 and A_3 received 2. Then the appropriate ARFF file would look like such:

```
@relation animal_example
@attribute "cat" { 0, 1 }
@attribute "dog" { 0, 1 }
@attribute "mouse" { 0, 1 }
@attribute "horse" { 0, 1 }
@attribute "rabbit" { 0, 1 }
@attribute "rabbit" { 0, 1 }
@attribute "monkey" { 0, 1 }
@attribute "classification" { 0, 1, 2 }

@data
1, 1, 1, 0, 0, 0, 1
0, 1, 0, 1, 1, 0, 0
0, 0, 0, 1, 0, 1, 2
```

Notice that we must determine the maximum mark in the training set in order to define the range for the "classification" attribute. Upon collating the attributes, it is beneficial to maintain a count of all attributes. Any attribute occurring only once is removed, as it is deemed as having insufficient impact on the classifier. This has the benefit of massively reducing the number of attributes, thus making training and classification considerably faster.

Training a classifier in Weka is near trivial using an ARFF file. We construct an untrained classifier, and an object representing the data in the ARFF file. After providing a reference to this data in the classifier, the training process is finally invoked².

² The trained classifier is required for future use, so we use Java's *FileOutputStream* class to write the classifier to a file. File representations of classifiers are (arbitrarily) given 'model' as a file extension.

There are a number of established classifiers that one may wish to use. Here is a brief description of the options available:

- Neural Network: This consists of a number of connected units known as 'neurones',
 analogous to the functioning of the brain. Each neurone takes the weighted sum of its
 inputs, and applies a function (usually either a threshold or sigmoid function) to produce its
 output. The network uses the training set and error minimisation via a process of 'back
 propagation' to iteratively adjust these weights.
- Decision Tree: This repeatedly partitions the training set on attributes, ordered by the 'information gain' (i.e. the extent of diversity in the proposed partitioning) each attribute provides. This results in a tree, with each node splitting into two given the presence of a certain feature, and the 'leaves' designating the resulting classification.
- Naive Bayesian Classifier: This uses simple probabilistic theory based on Bayes' Theorem. It
 calculates the probability of a classification (given the features) by using the probability of
 each feature given the classification and the prior probability of each classification, both of
 which are easily calculated. The classification which maximises this probability is selected. It
 assumes that each of the features are independent.
- Maximum entropy modelling: A framework for integrating information from many sources
 for classification. The idea is that it finds a model that satisfies the given constraints but does
 not go 'beyond the data', that is, it avoids a model not justified by the empirical evidence
 (i.e. the training set). Its advantage over the Naive Bayesian classifier is that it does not
 enforce independence assumptions with regard to the features.

Each of these classifiers has its own advantages and disadvantages. For the purposes of the evaluation process, the decision tree and naive Bayesian classifiers will be used, given their support by the Weka libraries.

Given our 4 stages of semantic parsing, featurisation, data preparation and training, we now have a complete system suitable for testing. However, there are some possible refinements we can make given the specific nature of the classification problem at hand. Two issues raised and explored are that of spelling correction, and a new form of classification that deals with the cumulative nature of exam marking.

6. Spelling Correction

Exam answers, typically those of a younger contiguate, must be expected to be filled with both grammatical and spelling errors. While the former are incredibly difficult to correct by autonomous means, it is possible to develop a system which corrects spelling errors with a good degree of accuracy.

One must initially question the importance of this correction in marking. Errors in inconsequential words which form the structure of the sentence will be of little concern to the examiner. However correcting such words would seem vital in order for the parser and boxer to establish a more accurate interpretation of the sentence, and to reduce the number of attributes in training. On the other hand, misspellings of keywords directly encapsulated by the mark scheme may heavily influence the examiner's marking. While the 'examiner's generosity' is possible to approximately model, it raises the issue that 'desirable spelling correction' is a fairly ambiguous and nondeterministic problem. This would suggest that a semi-autonomous approach should be adopted rather than a fully autonomous one. The system should produce a modifiable mapping defining how spelling corrections should be made, before mapping this (whenever the administrator chooses to do so) on the data.

An important observation to make is that a universal set of correct spellings is difficult to obtain. Many words are subject specific and thus will not be found in a standard dictionary. Therefore we must use a combination of the words in the data and words in a dictionary. A simplified version of the proposed correction algorithm can be defined as thus:

```
correction(word) = if correct(word) then word else arg max ( { score(x,word) | x \in suggestedCorrections(word) \} \cup { score(x,word) | x \in suggestedCorrections(word) } \cup { score(x,word) | x \in
```

where score(x,y) and score`(x,y) are two different scoring functions for dictionary and corpus comparison. We distinguish between the 2 functions since we wish to take into account the frequency of words in the corpus.

Dictionary Comparison

Upon obtaining a list of suggestions for an incorrectly spelled word, we obtain a similarity measure between the word w and the suggested correction s:

```
score(w,s) = \alpha.FirstLetterBonus(w,s) - \beta.LevenshteinDistance(w,s) - LetterDifference(w,s) - SizeDifference(w,s)
```

It was found $\alpha = 3$ and $\beta = 2$ produced the best results. A number of functions, of the type (String, String) $\mapsto N$, provide various string metrics:

| Name | Explanation | Motivation | Implementation |
|----------------------|---------------------------|---|---|
| Levenshtein Distance | deletion, or substitution | This provides an initial well-established comparison measure. | A dynamic programming algorithm using a (n + 1) × (m + 1) matrix, where n and m are the lengths of the two strings. |

| First Letter Bonus | Returns 1 if the first letter | Regardless of how badly | |
|--------------------|-------------------------------|-------------------------------|-------------------------------|
| | of both words is the same. | a word is spelled, | |
| | and 0 otherwise. | typically the first letter is | |
| | and o other wise. | correct. Thus from | |
| | | "recieve", it is more likely | |
| | | · · · | |
| | | that "receive" was | |
| | | intended than "believe", | |
| | | despite both having the | |
| | | same Levenshtein | |
| Law Diff | m 11 | distance. | |
| Letter Difference | | We favour corrections in | #[letterset(s) \cap |
| | | which the letters used is | letterset(w) ^r] + |
| | | more consistent. This for | #[letterset(s)' \cap |
| | are in one word but not | • | letterset(w) } |
| | the other. | correction from "doen" to | |
| | | "done" rather than the | |
| _ | | less probable "den". | |
| Size Difference | The difference in size | This gives more | # w - # s |
| | hetween w and s | preference to a letter | |
| | | pair swap' over an | |
| | | addition/deletion, and | |
| | | elíminates candidate | |
| | | corrections too different | |
| | | in size. | |

We keep the correction with the highest score, but delay adding this to the mapping given that this may be overwritten by a better match in the corpus.

Corpus Comparison

When comparing a misspelled word against the corpus (i.e. the exam answers), the assumption above that the words are correctly spelled can no longer be made; therefore the frequency of each term must be taken into account. We can augment the previous metric to provide a new one:

$$score'(w,s) = \mu \cdot log frequency(s) + score(w,s) - \lambda$$

In essence, we provide extra weighting if the word is commonly occurring, that is, we can be more confident of the word's 'correctness'. The constant λ ensures that it is not necessarily the case that score'(w,s) > score(w,s) (which would essentially discard dictionary comparison). We make a number of checks to prevent incorrect mappings:

- 1. Most obviously, we ignore the pair (w,s) if w is equal to s (preventing the identity mapping).
- 2. We ignore (w,s) if the Levenshtein Distance between them is greater than 5. This prevents a mapping to a very dissimilar word if it is the best that exists (i.e. it may be the case there is no mapping available at all).
- 3. We ignore (w,s) if the frequency of w is greater than the frequency of s in the corpus. This prevents mapping cycles, which would cause an infinite loop when attempting to map the corrections on the data.

Other Considerations

There are a few forms of errors that cannot be easily detected by the above methods. We can correct these accordingly:

- 1. Plurality errors: It is common for the plural form of a word to be misspelled. For example, "babies" is commonly written incorrectly as "babys", and "halves" as "halfs". In the above techniques, it is likely that an alternative word will be suggested, therefore we must compensate accordingly. We hypothetically replace common incorrect suffices with their correct equivalent, such as "ys" for "ies" and "fs" for "ves". If the new word is in the dictionary, we add this to the mapping and ignore any other correction methods.
- 2. Words augmented with apostrophes: Nouns can have "s" added to the end. The new string is not in the dictionary, despite this extension being perfectly valid. We check if the substring before the apostrophe is in the dictionary, and if so, regard the word as correct.
- 3. Word splitting: It is common for certain pairs of words to be incorrectly merged, for example "aswell", "alot" and "thankyou". To split these, we hypothesise a space insertion at each point in the string; if this forms two words both found in the dictionary, we add this to the mapping. This method can be problematic if not augmented, given that unlikely splits such as "aswell" into "a swell" can be made. This problem can be partially alleviated by favouring splits in which the bigram occurs most frequently in the corpus (i.e. "as well" is likely to occur more frequently than "a swell").

Mapping the Corrections

- 1--

The mappings of incorrect words to their (hopefully) correct forms are saved in a file. This allows the user to inspect the suggested mappings and either add their own or modify/correct the ones already present. The mapping is applied simply by searching for occurrences of each member of the mapping key in the data, and replacing it with its corresponding value.

A Demonstration

1 - 4

Here is a sample of corrections produced when dealing with a question on 'vasoconstriction':

| alot | a lot | hypothla |
|--------------|-------------|-----------|
| ammount | amount | insulate |
| aswell | as well | inturn |
| bodys | bodies | mantain |
| capilaries | capillaries | normaly |
| capilary | capillary | prents |
| capileries | capillaries | presonn |
| capillarites | capillaries | stoped |
| capillars | capillaries | suplie |
| capillarys | capillaries | transfer |
| capilleries | capillaries | traveling |
| capilliaries | capillaries | vascoco |
| capilliery | capillary | vascond |
| cappillaries | capillaries | vascons |
| cappillary | capillary | vasconti |
| cappilleries | capillaries | vasilaco |
| colser | colder | vasocon |
| dont | don't | vesseles |
| errect | erect | vessles |
| everytime | every time | |
| | | |

| hypothlamus | hγpothalamus |
|--------------------|-------------------|
| insulater | insulator |
| inturn | in turn |
| mantain | maintain |
| normaly | normally |
| prents | parents |
| presonne | person |
| stoped | stopped |
| suplie | supply |
| transfered | transferred |
| traveling | travelling |
| vascoconstrict | vasoconstrict |
| vascondriction | vasoconstriction |
| vasconstriction | vasoconstriction |
| vascontriction | vasoconstriction |
| vasilaconstriction | vaso constriction |
| vasoconstruction | vasoconstriction |
| vesseles | vessels |
| vessles | vessels |
| | |

7. Cumulative Classification

The approaches used thus far produced promising results despite the small training sets available. However, there is a significant conceptual flaw in the classification approach used, stemming from the fact that classifiers have no abstract notion of 'magnitude' in their outputted values. Marks assigned may as well be replaced with arbitrary letters or symbols, since the classifiers cannot distinguish that a mark of 2 is greater than a mark of 1. This leads to some unfortunate consequences:

- As the maximum mark becomes large, standard classifiers rapidly deteriorate in performance. Consider for example all answers which are awarded 5 out of 10 marks. These marks could have been awarded in ${}^{10}C_5 = 252$ ways. Clearly, even with a large training set, it would be impossible for classifiers to fully encapsulate 5 mark answers, when one considers that each of these 252 ways are based on multiple features, further increasing the uncertainty.
- We neglect dependency assumptions. If an answer is classified as 2 and satisfies some property A \times B, then clearly if two answers both classified as 1 have the properties A and B respectively satisfied, then they are strongly correlated. By ignoring these dependencies, we surrender a large quantity of inference that can potentially greatly aid in the training process.

A possible solution we propose is to introduce a notion of 'cumulative classification', with just one fundamental property:

If two sets of properties X and Y are satisfied resulting in classifications x and y respectively, then the overall resulting classification for $X \cup Y$ is x + y (given that X is not a subset of Y, and vice versa).

This is an intuitive principle, and encompasses the way in which an examiner would mark a paper. If an answer gets a mark for mentioning one point and a second for another, then one would expect this to contribute a total of 2 marks to their final mark. Standard machine learning classifiers cannot represent this concept.

The method proposed can be divided into 3 distinct stages:

- Segmentation: We divide the features into sets of features relating to each mark in the mark scheme. These sets may not necessarily be disjoint, neither must they span the whole set of features. We maintain a classifier for each of the marks, with the corresponding sets denoting each one's attributes.
- 2. *Training:* For each answer, we determine which of the marks are most likely to be satisfied and which aren't, and train each classifier accordingly.
- 3. Classification: We combine the outputs of these classifiers to provide an overall classification for the answer.

In summary, we are delegating the problem to each of the individual marks, combining their results to solve the overall classification task. The following sections deal with the motivation and implementation for each of the three stages.

Segmentation

The cumulative classifier is a composite of simpler classifiers managing the classification of each individual mark. Each of these classifiers requires a list of attributes, and intuitively one would expect this to incorporate features relating to this particular mark. There are two measures which we must consider to determine the accuracy of this algorithm. The first is precision; the proportion of features which appear in the 'optimum group'. The second is recall; the proportion of features from the 'optimum group' that appear.

Precision is important, when we consider that taking the entire set of features for all groups, which produces 100% recall but low precision, would result in poor classification, given that the system has no way to distinguish between marks. Recall is also important, given that neglecting important features produces a classifier that is not representative of the mark.

Some consideration should be given to the relationship between the sets of features. Should for example the sets be disjoint? This depends greatly on the mark scheme in question. Marks tend to be based on independent points to avoid ambiguity in marking, but it is very possible that two different marks will mention the same item or action. The sets need not provide full coverage of the entire feature set, when we consider that many points contained within some answer may be entirely irrelevant to the mark scheme, and thus not associated with any particular mark.

We construct 3 different methods that attempt to perform this segmentation. The first 2 of these use a 'co-occurrence' matrix, that is, a symmetric $n \times n$ matrix (n is the number of features) where each element $M_{ij} = M_{ji}$ is the number of times feature i occurs with feature j. This can be interpreted as a measure of the connection between 2 features. To compute this, we take a binary matrix A representing the training data, such that each row is an incidence vector representing one answer³. M is easily calculated:

$$M = A^T \cdot A$$

This follows from the fact M_{ij} takes the dot product of columns i and j (where such a vector is the incidence vector of a particular feature across all answers), equating to the number of times both features occur across all answers (i.e. if both features are present in a particular answer, their product is 1, contributing to the total count, otherwise their product is 0). We could feasibly use M^k for some k > 1 to improve the connectivity of features. If k = 2 for example, then M^2_{ij} conceptually indicates the extent to which feature i is connected to feature j via some intermediate feature; if for example "cat" occurs with "fur" in an answer and "fur" with "dog", but not "cat" with "dog", then M^2 allows us to make such a connection between "cat" and "dog". In practice this has an adverse effect for the type of data in question. We finally map the log function onto all elements of the

³ An initial assumption was that filtering the matrix to rows that received '1' as a classification may provide an optimisation. This is because each row would have present features mostly corresponding to one mark, thus increasing the probability that two features have a high co-occurrence due to them relating to the same mark. However, this reduced the training set to such an extent that this was not a viable approach.

matrix before normalising the columns⁴. This ensures that particular high co-occurrence values do not dominate and force other values towards 0 after normalisation.

Method 1: Standard K-means Clustering

K-means clustering is a commonly used method that partitions elements into disjoint sets (which span the entire set) using a standard algorithm. The resulting clusters have been partitioned in such a way that the mean distance from each vector to the centroid⁵ of its respective cluster is minimised. The process consists of simply clustering on the columns of M into k sets, where k is the maximum mark. Weka provides clustering facilities, thus the implementation involves producing a 'wrapper' which uses the matrix M to construct an input suitable for the clusterer, before returning a list of sets, where each set contains the indices of the features in the cluster.

The algorithm is fast and efficient, and provides moderate accuracy for most questions. However the disjoint set assumption is likely to cause problems for questions where high feature overlap occurs between marks.

Method 2: Iterative Clique Merging

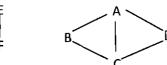
An alternative approach proposed here generates a number of seeds, where each seed is an island of closely connected features, before merging them into the required number of clusters.

Instead of normalising the columns in M as before, we scale all elements down so that the maximum element is 1. We define a threshold α (0 < α ≤ 1) such that if $M_{ij} \ge \alpha$, the bond between features i and j is considered to be 'strong', and 'weak' otherwise. This allows us to construct an undirected graph G, where we define the nodes to be the features, and the edges to be the 'strong bonds'.

To demonstrate this, take an example such that the co-occurrence matrix M is:

| | Α | В | С | D | E | F |
|-------------|-------------|---|---|---|---|---|
| Α | 1 1 1 | 1 | 1 | 1 | 0 | 0 |
| B C D | 1 | 1 | 1 | 0 | 0 | 0 |
| C | 1 1 1 | 1 | 1 | 1 | 0 | 0 |
| D | 1 | 0 | 1 | 1 | 0 | 0 |
| Ε | 0 | 0 | 0 | 0 | 1 | 1 |
| F | o | 0 | 0 | 0 | 1 | 1 |

with features A to F. The value of α is arbitrary here since all non-zero values are 1. This produces the graph:



⁴ More specifically, the function is f(x) = log(x+1). This ensures that 0 maps to 0, rather than produce an error.

⁵ The centroid of a cluster is the mean of its constituent vectors.

We start by identifying all cliques contained within the graph. A clique is a set of nodes such that there is an edge between every pair of nodes within it. From the graph, we can see that there are 3 cliques: {E, F}, {A, B, C} and {A, C, D}. To identify the cliques, we devise the following algorithm:

```
Let cliques = \emptyset
For (each row i of M)

If (cliques is empty) add new clique (i) to cliques

Else (

For (each clique q in cliques)

If all the bonds from i to each af the elements in q are strong, add i ta q.

Else if all the bonds from i to each of the elements in q are weak (and additionally, i was not added to any other clique in cliques) create a new clique (i) and add it to cliques.

Else (i.e. we have a mixture of weak and strong bonds) create a new clique:

q' = (i) \cup (j \mid strong(i, j), j \in q) and add it to cliques.
```

Presuming that *cliques* is not empty, we observe for each of the accumulated cliques which nodes the node *i* has edges to. Clearly if there are edges to all the nodes, the clique is preserved so we add the node *i* to the set. Similarly, if *i* has no edges to any element in the set of cliques (such as row E), we create a new clique (delayed until the end of iteration to ensure this condition is met). Lastly, if there are edges to some but not all the nodes in a clique (such as node D with {A,B,C}), we create a new clique containing the node itself and those nodes which it has an edge to. One might initially think this is a polynomial time algorithm⁶ given the nested for loops. However, the inner loop becomes larger as *cliques* expands, leading to an exponential algorithm.

If we follow this algorithm for the above example, on each iteration we obtain: [{A}], [{A,B}], [{A,B,C}], [{A,B,C}, {A,C,D}] (where the split occurs), [{A,B,C}, {A,C,D}, {E}], [{A,B,C}, {A,C,D}, {E,F}].

We desire only k groups however, so the sets above must be merged. Merging is performed by constructing an appropriate algorithm:

```
Let groups = cliques; Discard all g ∈ graups of size 1

While( |groups| > k)

Let bestGroup1 = Ø

Let bestGraup2 = Ø

For( each group g1 in groups)

For(each group g2 in groups)

Take the cross product of g1 and g2 (g1 x g2) and

find the average band strength over all pairs. If this value

exceeds the best so far, set bestGraup1 = g1, bestGroup2 = g2.

Add bestGroup1 ∪ bestGraup2 to groups.

Remove bestGroup1 and bestGroup2 from groups.
```

⁶ If this were the case, then P = NP given that a similar problem, the 'maximum clique' problem, appears as one of Richard Karp's original 21 problems shown NP-complete in his seminal 1972 paper "Reducibility Among Combinatorial Problems".

Suppose k = 2. Then just after one iteration of the algorithm, $\{A,B,C\}$ and $\{A,C,D\}$ will be merged to $\{A,B,C,D\}$, leaving this and $\{E,F\}$. Observing the original diagram, such group identification for this particular example appears sensible.

The fundamental flaw in this algorithm is that its unavoidable exponential complexity renders segmentation on hundreds of features (as would usually be expected) infeasible. Even if divide-and-conquer techniques could be used on the merging to produce an O(n.logn) algorithm, there exists tens of thousands of cliques produced from the previous process. Using singleton sets with each of the features as the seeds for the merging algorithm simply does not work as the resulting sets will be disjoint; in which case the standard clustering algorithm would be a superior alternative.

One should also note that a slight alteration in α can drastically alter the output of the algorithm. Clearly tweaking such a parameter dependent on the data set to improve results is undesirable.

Method 3: Manual specification

A non-autonomous alternative is to manually specify the groups each feature appears in. Clearly manually assigning features to groups removes the problems with the methods listed above, and any loss in accuracy is due only to human error. Given that a data set of 150 answers typically produces 800 features, the task is an arduous one and becomes impractical as the data set increases. A command line interface developed can alleviate this problem if the mark scheme is based on simple keywords (see *Implementation Notes*).

Training

The segmentation from the previous stage produces k classifiers with assigned attributes ready for training. In order to train multiple networks, we must determine a method of taking one instance and distributing this over the k classifiers.

Suppose an answer in the dataset obtained a mark of j out of k ($0 \le j \le k$). Each classifier can only produce a classification of 0 or 1 (indicating whether the mark was obtained or not), so we expect to train j of these classifiers with 1 (where the instance is restricted to the features the feature set corresponding to that classifier contains), and the remaining k-j with 0.

To determine which of these j classifiers are trained with 1, we rank the k groups in order of cosine similarity between the cluster's vector (with 1s for attributes that occur in the cluster, and 0s for those that don't) and the instance vector, from highest to lowest. In practical terms, this gives a relative measure of the magnitude of the instance 'matching' each cluster. Once this ranking is achieved, we simply take the top j elements from the list as the clusters that 'received' the mark.

The cosine similarity is a measure of the similarity of two vectors by calculating the angle between them. For two vectors A and B, it is defined to be:

$$\cos\theta = \frac{A \cdot B}{|A||B|}$$

We can ignore the Euclidean length of the vector corresponding to the answer to classify, since this will remain constant as we vary the cluster vector. Thus the measure is simply:

$\frac{cluster.answer}{|cluster|}$

Classification

For a new instance, we simply restrict the answer vector to each of the clusters, classify each instance, and sum over the classifications. This is analogous to a marker determining which marks in the mark scheme were awarded, before totaling their score.

Extending to alternative types of mark scheme requires little change. Take for example a scheme of the form "Award 1 mark for any of the k points, maximum n". The number of clusters remains the same, but in classification we take *min(mark, n)* so that the maximum mark is not exceeded.

A worked example

Suppose a mark scheme consisted of 2 marks, with the following criteria:

- 1) Mention of a 'dog'.
- 2) Mention of the act of 'walking' or 'exercise'.

Suppose also the following training answers were supplied, their classification denoted in brackets:

- A. The big dog started barking. (1)
- B. The dog buried his bone. (1)
- C. Adam enjoys walking and exercise. (1)
- D. The dog enjoyed exercising. (2)
- E. The big lecturer scared the timid student. (0)

Each of the answers would be semantically analysed and deconstructed into its constituent features. For the sake of simplicity, we extract only 4: 'dog', 'big', 'walk' and 'exercise'. We could therefore generate the following incidence matrix A for the above answers:

| | dog | big | walk | exercise |
|---|-----|-----|------|----------|
| Α | 1 | 1 | 0 | 0 |
| В | 1 | 0 | 0 | 0 |
| С | 0 | 0 | 1 | 1 |
| D | 1 | 0 | 0 | 1 |
| E | 0 | 1 _ | 0 | 0 |

Computing $M = A^T$. A, taking the log of each value, and normalising the columns, we obtain:

| | dog | big | walk | exercise |
|----------|-----|-------|------|----------|
| dog | 0.5 | 0.387 | 0 | 0.279 |
| big | 1 | 0.613 | 0 | 0 |
| walk | 0 | 0 | 0.5 | 0.279 |
| exercise | 1 | 0 | 0.5 | 0.442 |

By inspection of the non-zero values, it is clear that the two groups obtained by clustering are $G_1 = \{$ dog, big $\}$ and $G_2 = \{$ walk, exercise $\}$. This is consistent with the original mark scheme. We create a standard classifier for each of these groups (say C_1 and C_2), and now use the rows of the matrix A to

train these. We define an ordered list L_R to be the groups in descending order of their similarity with the row R.

The incidence vectors for the 2 groups are (1, 1, 0, 0) and (0, 0, 1, 1) respectively. Since the Euclidean distance of these are the same, for this example we can ignore these from the cosine measure and simply use the dot product instead.

For the first row A, we have the vector (1, 1, 0, 0). We take its dot product with each of the group incidence vectors:

$$(1, 1, 0, 0) \cdot (1, 1, 0, 0) = 2$$

 $(1, 1, 0, 0) \cdot (0, 0, 1, 1) = 0$

Since 2 > 0, this gives us $L_A = [G_1, G_2]$. For this first row, 1 out of 2 marks were awarded, so we train C_1 (corresponding to G_1 , the first element of the list) with the row (restricted to the features it contains, i.e. (1, 1)) and the classification '1', and train C_2 with the row (restricted to (0, 0)) and the classification '0'.

Using this method for each row, we train each classifier 5 times.

| | 1. 1. | | gar in | C_1 | | ·C ₂ | |
|-----|----------------------|----------------------|--------|-------------------|-------|-------------------|-------|
| Row | Row . G ₁ | Row . G ₂ | Marks | Training instance | Class | Training instance | Class |
| Α | 2 | 0 | 1 | (1, 1) | 1 | (0, 0) | 0 |
| В | 1 | 0 | 1 | (1, 0) | 1 | (0, 0) | 0 |
| С | 0 | 2 | 1 | (0, 0) | 0 | (1, 1) | 1 |
| D | 1 | 1 | 2 | (1, 0) | 1 | (0, 1) | 1 |
| E | 1 | 0 | 0 | (0, 1) | 0 | (0, 0) | 0 |

Now that training is complete, classifying a new answer is a simple process. For example, the answer "The big hamster had some exercise on his wheel" is converted to the vector (0, 1, 0, 1) using the 4 features. We take the vector (0, 1) corresponding to the group G_1 and put it through C_1 . The classifier hopefully should return '0' given that a dog isn't mentioned and the 'big' adjective is irrelevant. Next we take the vector (0, 1) corresponding to the group G_2 and put it through C_2 . The classifier should return '1' given that 'exercise' is mentioned. 0 + 1 = 1, therefore the final classification for this answer is 1.

Pre-Testing Analysis

Before this system is tested, it is beneficial to establish a hypothetical upper bound on accuracy compared with conventional classifiers. While cumulative classification has some merits, its cumulative nature leads unfortunately to cumulative error.

Consider the case where the maximum mark is 1. Since partitioning the feature set into 1 leaves it unaffected, the algorithm behaves the same as a normal classifier. Suppose that we assign a probability p that the classification for an instance is correct, and for simplification, presume this is the same probability as classifying any mark in a mark scheme correctly. If a mark of k out of n was obtained, we obtain the following probabilistic upper bound for a correct overall classification being received:

P(correct) =
$$\sum_{r=0}^{\min(k, n-k)} {k \choose r} {n-k \choose r} p^{n-2r} (1-p)^{2r}$$

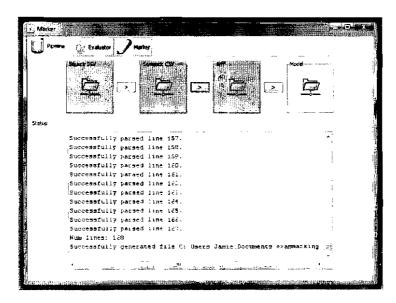
This upper bound could only theoretically be reached if the segmentation is optimal, and if in training we always identify the correct clusters for which a mark was awarded. This expression seems complex, but is a simple modification of a standard binomial distribution. To understand it, consider firstly the probability of the examiner classifying all marks correctly. This is simply p^n . Now consider that the examiner accidently assigns 2 marks incorrectly, such that for one mark it was wrongly identified as incorrect, and another wrongly identified as correct. The total classification would remain the same, since we are not changing the total number of marks. There are ${}^kC_1 = k$ ways of picking a 1 to switch to a 0, and ${}^{(n-k)}C_1 = (n-k)$ ways of picking a 0 to switch to a 1. (n-2) marks are assigned correctly with probability p (giving p^{n-2r}) and 2 marks are assigned incorrectly with probability (1-p). We generalize this to 2r mistakes where r is the number of incorrect pairs, and sum over all the possible number of pairs of mistake. There is a maximum of min(k,n-k) mistakes, since we are limited by the number of 0s we can pair with 1s, or 1s we can pair with 0s.

To assess the practical impact of this, take p to be a sensible value of 0.85. Presuming a random distribution of marks out of 4, the average upper bound for the probability of correct classification is 0.55462. Given that errors in clustering and training are likely to lower the accuracy, this upper bound is bad news. However, accuracy is not the only method of evaluation, and a measure known as 'disposition' is introduced later to justify the usefulness of this classification technique.

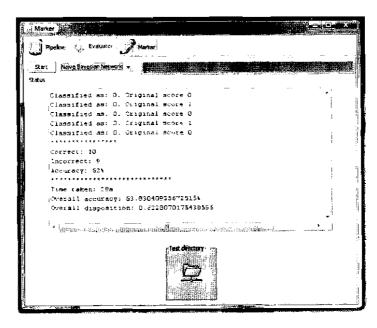
8. Graphical User Interface

Currently all methods and operations discussed previously are only accessible via the command line. When dealing with numerous files across the conversion pipeline, this can be somewhat of a cumbersome process. Therefore, a simple GUI (Graphical User Interface) was developed and used to facilitate such operations.

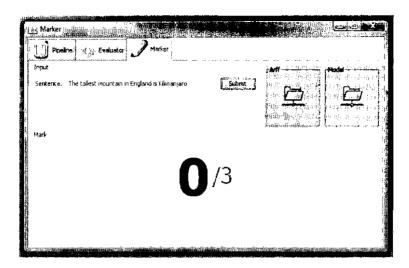
The GUI consists of 3 tabbed panels. The first allows conversion between various file formats across the overall pipeline. The first button takes a CSV of English answers and parses them to produce a CSV of DRS expressions. The second button uses this and produces an ARFF, and finally the last button uses the ARFF to train a classifier (prompting the user to select between the different classification methods).



A second panel allows the user to test the performance of the system for a given dataset. The user simply selects their desired training type and the folder containing the input data (which must contain a file named 'answers.csv'), and the results are printed.

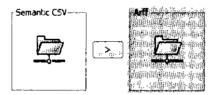


Finally, a third panel allows individual answers to be classified given a trained classifier.



There are two features of the GUI worth drawing attention to:

• Drag & Drop file selection box: The class FileDragPanel is simply a component, and allows the user to specify a file or directory quickly. It has 'drag and drop' functionality, allowing the user to drag a file or directory into the box to make the selection. Alternatively (given that the file to specify may not yet exist if it is the output of an operation) the user can click the panel to open a standard browse dialog. The constructor takes a file type (enforced when browsing) and a name, which is displayed at the top of the panel.



Loading bar: The evaluation process can be long, and thus it is desirable to give the user a rough indication of the time remaining. Progress bars are easy to create and update in the Swing framework. However, an issue arises of GUI components becoming 'frozen' until the entire evaluation process is complete, resulting in a jump from 0% to 100% as it finishes. The solution is to run the evaluation process 'in the background'. The abstract class SwingWorker<Void, Void> has a method doInBackground(); by implementing this, we ensure that updates to the GUI during an operation are administered. The progress is merely an approximation, with 0-20% progress for parsing, 20-40% for generating the ARFF files, 40-80% for training and 80-100% for testing.



The GUI used the Swing toolkit available to Java. The 'Look & Feel' (that is, the interface style) is set to match that of the operation system; as one may have observed, all screenshots were taken while using Windows Vista.

9. Evaluation

The form of evaluation used employs a technique known as "10-fold cross-validation", commonly used in Computation Linguistic systems. This splits the training set into 10 segments, trains a classifier on 9 of the segments and classifies the remaining 1 segment. The results are averaged over the 10 possible testing configurations.

Each answer in the test segment is classified, and is compared to the actual classification from the data. A measure of the system's performance is obtained, encompassing two properties:

- Accuracy: The percentage of answers that were classified correctly.
- Disposition: The average displacement of the estimated marks to the actual marks. This is potentially a more useful measure for higher mark questions, since for a mark of 4 out of 4, an estimation of 3 is superior to an estimation of 0.

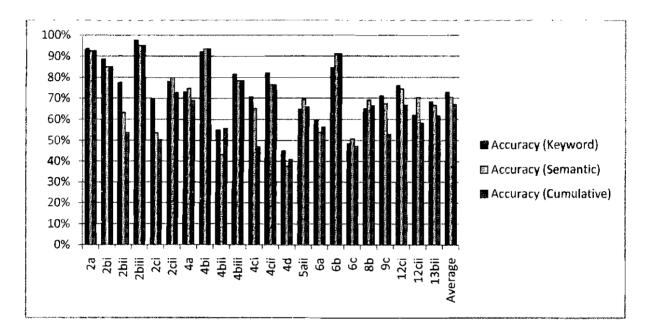
The Naive Bayesian Network was chosen as the classifier (and for the cumulative classifier as the sub-classifiers). The decision tree classifier gave on average approximately 5% worse performance, and thus was ignored. Results were obtained for the 3 different approaches: the keyword approach, the semantic approach using the parser, and the cumulative classifier described in Section 7.

Results

| | | | Keyword / | Keyword Approach | | Semantic Approach | | ti v e er |
|----------------|-------------|----------------|-----------|------------------|-----------------|----------------------|----------------|---------------------|
| Q | Max mark | Num answers | Accuracy | Disposition | Acc. | Dis. | Acc. | Dis. |
| 2a | 1 | 200 | 94% | 0.065 | 92.50% | 0.075 | 92.50% | 0.075 |
| 2bi | 1 | 192 | 89% | 0.115 | 84.87% | 0.151 | 84.87% | 0.151 |
| 2bii | 2. | 190 | 78% | 0.225 | 63.16% | 0.374 | 53.68% | 0.489 |
| 2biii | 1 | 200 | 98% | 0.025 | 95% | 0.05 | 95% | 0.05 |
| 2ci | 2 | 189 | 70% | 0.335 | 53.48% | 0.518 | 50.32% | 0.534 |
| 2cii | 2 | 193 | 78% | 0.220 | 79.92% | 0.201 | 72.68% | 0.278 |
| 4a | 2 | 162 | 73% | 0.306 | 7 4 74 % | 0.296 | 69.01% | 0.322 |
| 4bi | 1 | 195 | 92% | 0.080 | 93.34% | 0.067 | 93.34% | 0.067 |
| 4bii | 2 | 192 | 55% | 0.465 | 43.29% | 0.687 | 55.76% | 0.504 |
| 4biii | 1 | 83 | 81% | 0.186 | 78.3 3 % | 0.217 | 78.33% | 0.217 |
| 4ci | 2 | 195 | 71% | 0,305 | 65.13% | 0.379 | 46.92% | 0.612 |
| 4cii | 1 | 193 | 82% | 0.180 | 76.29% | 0,237 | 76.29% | 0,237 |
| 4d | 3 | 163 | 45% | 0.625 | 37.65% | 0.708 | 40.88% | 0.77 |
| 5aii | 2 | 170 | 65% | 0.370 | 69.41% | 0 .3 29 | 65.88% | 0.37 |
| 6a | 3 | 186 | 60% | 0.490 | 53.83% | 0.623 | 56.37% | 0.506 |
| 6b | 1 | 201 | 85% | 0. 155 | 91.00% | 0.09 | 91.00% | 0.09 |
| 6 c | 4 | 139 | 48.5% | 0.644 | 50.71% | 0.707 | 47.14% | 0.636 |
| 8b | 2 | 186 | 65% | 0.366 | 69. 06 % | 0.332 | 6 6.26% | 0.382 |
| 9c | 2 | 182 | 71% | 0.300 | 67.35% | 0.332 | 52.75% | 0.478 |
| 12ci | 2 | 168 | 76% | 0.240 | 74.41% | 0.256 | 66.73% | 0.339 |

| 12cii | 2 | 158 | 62% | 0.415 | 70.29% | 0.316 | 58.17% | 0.463 |
|-------|---|-----|-----|-------|--------|-------|--------|-------|
| 13bii | 4 | 119 | 68% | 0.522 | 66.67% | 0.567 | 61.67% | 0.558 |
| Avg | | | 73% | 0.309 | 70.47% | 0.340 | 67.07% | 0.370 |

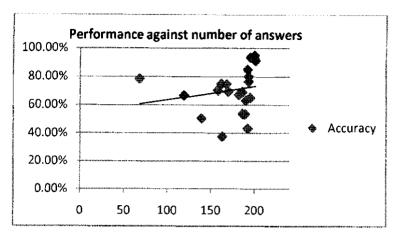
| | Keyword Approach | Semantic Approach | Cumulative Classifier |
|------|------------------|-------------------|------------------------------|
| Mark | Average Accuracy | Average Accuracy | Average Accuracy |
| 1 | 88.71% | 87.33% | 87.33% |
| 2 | 69.45% | 66.39% | 5 9 . 8 3% |
| 3 | 52.50% | 45.74% | 48.63% |
| 4 | 58.25% | 58.69% | 54.41% |

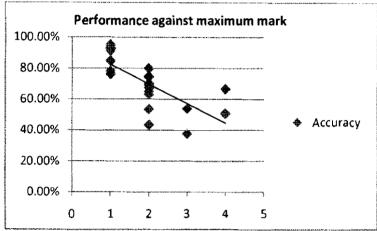


The first noticeable property of the above graph is that the semantic approach was marginally worse than the keyword approach. This is initially surprising, given that one would expect the set of features produced from the keyword approach to be a subset of that for the semantic approach, and thus the performance at least as good. There is a variety of explanations for this decrease in performance:

- Some variation of performance is expected given that the input data used was different.
 While the keyword approach used all input data, some badly spelled/ungrammatical answers were rejected by the parser, and thus discarded from the training data.
- 2. The parser had the effect of transforming some of the words, such as changing plurality or removing tense, and other forms of 'noise'.
- 3. Similarly, the parser is not 100% accurate, and therefore some answers will have an incorrect semantic representation.

More generally, the results show impressive performance for 1 mark questions, but a considerable decline in accuracy as the mark increased. The cumulative classifier gave relatively poorer performance to the semantic approach for 2 mark questions, but better performance for a few higher mark questions. Accuracy had some correlation with the size of the data set, albeit small.





The disposition data did not yield any significant trends across the 2 classifiers, other than being approximately proportional to the accuracy. There was only one dataset in which the accuracy was lower for the cumulative classifier but had a higher disposition. This is likely to be because the modal mark for each dataset tended to be 0; thus if a 4 out of 4 mark answer is classified (incorrectly) by the cumulative technique as 3, and as 0 by a standard classifier, the low frequency of such high-mark answers renders such improvement (i.e. a classification closer to the actual mark) insignificant.

For a given dataset, the wildly variable accuracies for each segment in the 10-fold cross-validation are also indicative of inadequate training data. If a large dataset was used, one would expect these accuracies to be roughly consistent. It should be noted that the parser was not always successful in producing a parse, particularly for long answers. This is unfortunately unavoidable, given the poor quality of grammar and spelling in many of the answers.

Cumulative Classifier

The poorer than expected results for the cumulative classifier can partially be explained by the often poor datasets. As identified above, the modal mark was often 0, and for higher mark questions, full-mark or near full-mark answers were very infrequent (in some cases appearing only once or twice). This, coupled with the low amount of training data, means that many of the individual mark subclassifiers predominantly receive a 0 as a classification in the training data.

It is curious that questions 4d and 6a had very contrasting performance despite both having the same maximum mark. Inspecting the mark scheme, this is likely to be because there was less overlap of features associated with marks in 6a (a question on plants, in which the marks were fairly

independent concepts) than there was with 4d (a question on vasoconstriction, in which 2 of the 3 marks made reference to blood). This strongly suggests that the cumulative classifier performs better when the clustering is less ambiguous.

Spelling Correction

| Question | Accuracy | Change | _ |
|----------|----------|----------------|---|
| 4a | 73.30% | -1.44% | _ |
| 5aii | 68.82% | -0.59% | |
| 6a | 52.78% | -1 .05% | |
| 6b | 88.05% | -2.95% | |
| 6c | 50.90% | 0.19% | |
| 8b | 68.54% | -0.52% | |
| 9c | 66.76% | -0.59% | |
| 12ci | 72.46% | -1.95% | |
| 12cii | 68.13% | -2.16% | |
| 13bii | 67.69% | 1.02% | |
| Average: | | -1.00% | |

The results show that spelling correction yields no improvement on accuracy. There are a number of reasons why there may have been a slight reduction in performance:

- The correction of words integral to the grammatical structure of the sentence led to less answers being rejected by the parser. This combined with the other alterations to the dataset leads to a minor change in accuracy, either positive or negative.
- The spelling corrections may be too generous that is, an incoherent answer with key terms
 misspelled may be penalised by the examiner. This results in some classifications in the
 training set being too high.

10. Conclusion

The results are a mixture of encouragements and discouragements. While the spell correction and cumulative classifier enhancements failed to provide an increase in performance, accuracy for many of the data sets was high. However, there was a disappointing lack of improvement in performance over the 'bag of words' approach, even if this performance was significantly superior to that of the 'k nearest neighbour' approach in [Pulman 06]. Unless there were fundamental inadequacies in the method of featurisation used, it is clear that semantic information from the sentences bears little or no influence on classification on the type of data used. This is perhaps of little surprise; while theoretically semantics are of paramount value, in the context of factual exam questions its use is limited. Typically one would not expect examinees to write incoherent answers containing the desired 'keywords' in an attempt to fool the system – instead the inclusion of such keywords in an answer is a strong reflection that the student has understood the topic at hand.

Despite this, the very high accuracies in some test segments suggest that accuracies could be significantly improved by datasets of a larger size, and thus one could deem the results 'inconclusive' given the inadequacy of the data.

Ultimately, the problem with a machine learning approach to automated exam marking is the existence of a no-win situation. A small training set leads to poor results, but a large training set is counterproductive for the classification problem at hand; each classifier can only be used for a specific question, and thus unless the total number of exam papers is very large, manually marking a large subset of these is not economical.

The conclusion is that the approach of manually specifying the mark scheme required (via the patterns previously explained in the introduction) is superior to attempting to learn such patterns. The cumulative classifier may aid in identifying possible groups of features for each mark, but given 100% accuracy in the segmentation algorithm is unrealistic, this leads to features occurring in incorrect groups, and thus renders such information limited in use. The cumulative classifier does however provide some 'feedback' where traditional classifiers would not, in that the classifier indicates which of marks in the classification were and were not awarded.

There are number of extensions that may be made to this research. A more accurate segmentation algorithm that doesn't rely on the co-occurrence matrix could be investigated. Relying solely on the co-occurrence matrix is possibly naive, as one cannot assume that two features occurring frequently together necessarily relate to the same mark (and as discussed, limiting the dataset to 1-mark answers is too restrictive). One might also propose that the system may effectively be used in another domain where one large dataset is used to classify instead of multiple smaller ones. Other optimisations could also be investigated, such as 'stemming', which merges classes of words containing variants of the same word (e.g. "fertilise", "fertilised" and "fertilisation") into a single representation.

11. Implementation notes

All coding was written in Java (1.6).

a) Parsing

The WebsiteParser class used for 'online parsing' simply inserts the English sentence into the URL (acting as a parameter), obtains the resulting page (using Java's InputStreamReader on the URL as a stream), and extracts the DRS expression from it.⁷

ServerThread is a class used to manage a particular connection. As an extension of Java's Thread class, such managers can be run concurrently to service multiple connections simultaneously.

Instructions to the SOAP server and client can be made via the command-line. We can execute such instructions from Java by constructing a *Process* object. This can be obtained by *Runtime.getRuntime(j.exec(command)*. The *Process* class usefully has a method *waitFor()*, which causes the method to halt until the scheduler signals completion of the process. One might wonder why inelegant file writing and reading were used in *ServerThread* instead of piping. This is because the command-line functionality is very limited in Java, and thus can only execute basic commands.

The following command-line prompts are used in implementing the server:

- soap_server --server localhost:9000 --candc ../candc/candc-1.00/models/boxer
 is used to start the server on the local machine (using an arbitrary port), specifying the
 Boxer data as the parsing model to use. Since there is no way of determining whether
 the parser has fully loaded, we wait a fixed amount of time (20 seconds) before
 continuing.
- soap_client --input inputFile --output parsedFile --url http://localhost:9000
 This command instructs the SOAP client to communicate with the SOAP server. The parsed output (in CCG form) is outputted to a file. The port must be the same as that specified in initialising the SOAP server.
- boxer --input parsedFile --output finalOutputFile --box true -flat
 This takes the parsed file and uses Boxer to produce the DRS expression. We extract this
 from the output file and write this to the output stream of the server.

b) Implementing φ

Implementing the ϕ function involves 2 stages. The first is to convert the DRS string obtained from parsing into some object oriented form. The second takes this construction to generate the desired atoms. The DRSreader class in the 'parser' package is used to execute this first stage, producing a SemanticModel object. A SemanticModel represents all the variables, predicates and relations (including equality, etc.) within the DRS.

To do this conversion, we establish an abstract class Expr. This has 2 unimplemented methods:

⁷ Such a technique of "web-scraping" is generally discouraged for large amounts of data, and therefore *WebsiteParser* can only be used to obtain parses for individual sentences (i.e. for classification rather than training purposes).

- int parse(List tokens, Integer index): Takes a list of tokens (where each token is a word or character of the DRS string) and the point to which we have currently scanned. This parses the tokens in a way relevant to the part of the DRS tree we are currently in, and builds up the internal structure of this object. The integer returned is the new token position after parsing.
- void process(SemanticModel model): Used for the second phase. It takes the structure of the object accumulated from parsing, and updates the SemanticModel.

We use an abstract class as opposed to an interface so that we can define a method *simpleParse*, which parses a single token (containing no information, such as parenthesis) and produces an error if the token is not as expected.

Clearly we must first convert the DRS string to tokenised form. A method *List<Object> tokenize(String str)* in *DRSreader* performs this, using Java's *StringTokenizer* class. To then initialise the parsing process, we just then construct a new *Drs* object (representing the root of the new tree being constructed), and then execute its parse method with these tokens. One might wonder why a *List* of tokens is used and not a more intuitive data structure such as an *Iterator*, where once a token is consumed, it is not seen again. The reason is that sometimes we wish to 'peek' at the current token without consuming it, in scenarios where there are multiple types of children that a node could have.

Now that a tree structure of the DRS has been obtained, we use it to construct *SemanticModel*. This object contains the following items:

- A map of variable names to SemonticVar, where a SemanticVar represents a variable in the DRS. Each has a reference to identify it (e.g. x1), a name of the object/event it is representing (e.g. dog), and a list of attributes (e.g. [hairy, large]). It has methods to set its name/add an attribute⁸, set its cardinality, and to obtain a list of its semantic atoms (as defined earlier).
- A list of SemanticRel, representing relations in the DRS. All relations are binary, therefore we
 have a reference to the left and right variables (we use their name rather than a pointer to
 the variable itself), and the predicate being applied. It similarly has a method to generate a
 list of semantic atoms as defined in Section 4.

SemanticModel maintains all other relations, that is, equalities and implications, and has methods to generate their atoms appropriately. It is possible to have nested SemanticModels, in the case when the DRS expression uses the 'Prop' (i.e. proposition) construction. When processing this, we construct a new child SemanticModel, set its parent to the current model, and redirect the processing of the sub-expression to the child model. Therefore when looking up a SemanticVar for a given name (a process required by SemanticRel), if it is not contained within the local context, we recursively search its ancestors.

The KeywordIdentifier class can be used to display the list of atoms for a DRS expression. While not used in the overall pipeline, it is useful for analytical purposes.

⁸ We can determine an entity's type by its 'posType' value. If we have hairy(x1) and hairy has the type 'a' (indicating an attribute) we add it to the list. If we have dog(x1) and dog has the type 'n' (for noun), we set the variable's name.

c) Dictionary Methods

The class *SpellUtils* provides access to methods associated with misspellings, such as determining whether a word is incorrect misspelled, and if so, producing a list of suggestions. An external library "Swabunga Spell Engine" was used, which provides such methods.

d) Cumulative Classification

Manual Mark Group Specification

A marklist file simply contains rows of the form "feature -> [list of groups]". If the directory the classifier model is being outputted to contains a file named 'main.marklist', then this will override any other segmentation method used. A class MarkListModifier provides a command-line interface to reduce the hugely time-consuming burden of manually assigning groups, with commands such as "CONTAINS ward group" to add any feature containing 'word' to the specified group.

Training

A class *ClusterUtils* provides methods to perform this ranking and converting vectors to their appropriate form:

List<Integer> getClustersBySimilarity(double[] row, List<Set<Integer>> clusters) takes the row and clusters and produces a ranked list of cluster indices using the cosine measure.

double[] getClusterIncidenceVector(Set<Integer> clusters, int numFeatures) puts the cluster in a vector form for use in the above method.

Lastly, double[] getRowRestrictedToCluster(Set<Integer> cluster, double[] row) removes values from the vector that correspond to features not in the cluster. This is the 'instance' used in training each classifier.

12. Bibliography

[PULMAN 05] S. G. Pulman and J. Z. Sukkarieh (2005): Automatic Short Answer Marking.

[PULMAN 06] Stephen G. Pulman, Nicholas J. Raikes and Jana Z. Sukkarieh (not yet published): Customising an automated marking system.

[SHERMIS 03] Mark D. Shermis, Jill Burstein, (2003): Automated Essay Scoring: A Cross-disciplinary Perspective.

[CLARK 07] James R. Curran, Stephen Clark, and Johan Bos (2007): Linguistically Motivated Large-Scale NLP with C&C and Boxer. Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo), pp.29-32.

Artificial examiners put to the test: http://news.bbc.co.uk/1/hi/technology/6961088.stm

Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). lan H. Witten, Eibe Frank (documentation for the Weka tools).

13. Acknowledgements

I would like to thank my project supervisor Professor Stephen Pulman for his helpful comments and pointing me towards useful sources and tools. I would also like to thank the OCR Exam Board for their provision of training data.

schemagenerator/ArffReader.java

```
€3:
 1: package schemagemerator;
                                                                                           6.4 :
                                                                                                           br.:cadline(); / @datu
3: import java.u. Bufine devater:
                                                                                           55:
 4: import fava.id.FileMeade:;
                                                                                           ée:
                                                                                                              Now aread the data.
                                                                                           :
 5: import have in Serial zamle:
                                                                                           ۶B:
 6: import hava.ut.1.linkodilish;
                                                                                                           int numbews = 0:
 /: import mava.um.i.list;
                                                                                           69:
                                                                                                           mas*<br/>Masrkboudde>> vals = new linked, is*<br/>dis*<br/><br/>boto e>> %;
                                                                                           70:
 f: import java.util.Map;
                                                                                                           while br. ready(1)
 9: import nava.util.Set;
                                                                                            12:
i0: import java.uril.TroeMap;
11:
                                                                                            /3:
                                                                                                                String[] lineFarts = pr. eail.net,.split(", ");
12: import woka.d re.Matrix;
                                                                                            16:
                                                                                                                list<.buble> row | new linkedList<br/>%louble> ();
13.5
                                                                                            5:
                                                                                                                numRows.+:
14: /**
                                                                                            -. 4.
                                                                                                                for(int y=0; i<(lineParts.gength-i); '++)</pre>
15: * This class provides access to the information contained
                                                                                            77:
                                                                                            78:
16: * within an Arff file, such as the number of features, each
                                                                                                                   Double va. = new Pouble(lineParts 11);
17: * of their counts, and a matrix representing the data.
                                                                                           79:
                                                                                                                   // If necessary, we increment the count for that feature.
18: * @author Jamie
                                                                                           80.
                                                                                                                   if (val -=1.0)
                                                                                           ც₁:
19: *
                                                                                                                   if(featureCounts.containsRey(1))
                                                                                           82:
20: +/
                                                                                                                            featureCounts.put(1, featureCounts.get(1)+1);
21: public class ArffResder implements Serializable
                                                                                           გვ:
                                                                                                                         else featureCounts.put(1, !);
                                                                                           84:
22: f
                                                                                           85:
23:
                                                                                                                   row.add(val);
     Matrix matrix:
      // The classifications as defined by the Arff file.
                                                                                           86:
24:
25:
      list<Integer> classifiedAs = new LinkedList<Integer>();
                                                                                           87:
                                                                                                                vais.add(row);
      // The total number of times each of the features (represented
                                                                                           88:
₹.
                                                                                           89:
27:
      // by its index) occurs, that is, the summation of the column.
                                                                                                                int mark = new Integer(lineParts[]ineParts.length-| );
                                                                                           90:
28:
      Mapkinteger, integer> featureCounts - new TreeMapkInteger, integer>();
                                                                                                                classifiedAs.add(mark);
29:
                                                                                           99:
                                                                                                                 if(mark>maxMark)maxMark : park;
      String fileName:
                                                                                           92:
30:
                                                                                           23:
31:
       int maxMark 0:
                                                                                           94:
32:
       List<String> attricutes = new Linkedlist<String>(;
                                                                                                           Matrix mx = new Matrix(numRows, numCo.s);
                                                                                           95:
33:
       public ArfiReader (String fileName)
                                                                                           96:
                                                                                                           for(int i=0; i<numRows; t++)</pre>
34:
                                                                                           07.
35:
            this.fileName tileName;
                                                                                                                for(int j=0; j<nurCols; j++)</pre>
                                                                                           98:
36:
            read():
37:
                                                                                           99:
                                                                                                                   mx.setblement(1,,,vals.get(1).det(1);;
38:
                                                                                          100:
                                                                                          131:
39.
                                                                                          102:
40:
                                                                                                           br.close():
41:
        * Reads the Arft file. Executed by the constructor.
                                                                                          103:
42:
                                                                                          041
                                                                                                           mat: x = ⊤x;
                                                                                          108:
43:
       public void read()
44:
                                                                                          1065
                                                                                                        1 catch (Exception e) {
45:
                                                                                          107:
                                                                                                           throw new Firor(e);
            try
                                                                                          108:
46:
                                                                                           , :
47:
               // Get attributes and read up to "#data"
               BufferedWeader by new BufferedReader(new F;leReader(fileWare));
                                                                                          110:
48:
                                                                                          111:
49:
               boolean reagric true;
                                                                                          112:
52:
               int manGals=0;
51:
               while 'br.ready()&&reading)
                                                                                          173:
                                                                                                   * Sets the prescribed mark for an answer with index i.
                                                                                          114:
52:
                                                                                          115:
                                                                                                   public Integer gotActualMark(Integer i)
53:
                    String line = pr.readLine();
                    if attricions.srze \left\( != 9&&line.redexOf("@attribute") < 0 \right\)</pre>
                                                                                          ា16:
                                                                                                       return of assifiedAs.get(1);
54:
F. - -
                       realing = false;
                                                                                          118:
                    else if (line.indext): ("@attribute" -== e) (
56:
                                                                                                   public Integer getFeatureCount(Integer i)
                                                                                          1100
                       String | parts | line.spl t." " :
37:
                                                                                                        return featureCounts.get(1);
                       if liparta () ad a's "score" (
58:
                                                                                          -20:
                            estiributes.edi(parks,[].rep aceAll("\"", "",);
5.83
                                                                                           ī:
                                                                                                   public Marrux us Vatrux.
60:
                            កលាលវិសាវិទា 🖅
                                                                                           2.3
                                                                                                       return fair.x;
                                                                                           12:11
61:
                                                                                          124:
€2:
```

schemagenerator/ArffReader.java

```
125:
         * @deprocated
136:
         * The morfood returns a matrix consisting only of rows with the Jiven matk
7:

    Plassification. Initially in training, only rows with a disself ratio.

120:
         * of I were taken. The logic was that such answers would contain toathies
120:
         I only relating to predominantly one mark, and thus ident fyind groups
137:
         * would be more accurate. However in practice this did seem to be the
131:
         \star case, and had the drawback of discarding the majority of the data set.
132:
133:
        public Matrix retMatrix(int mark)
34:
35:
136:
             int oils mairix, numColumns();
137:
             int numRowsWithMark = 0;
138:
             for(int i 0; i<classifiedAs.size(); i++)</pre>
139:
                 if (class: fledAs.get(i) == mark) numRowsWithMark++;
40:
141 -
             Matrix matrixForMark = new Matrix(numRowsWithMark,cols);
14.5:
143:
             int currentRow = 0:
144:
             for(int i 0; i<matrix.numRows(); i++)</pre>
145:
146:
                 if(classifiedAs.get(1).intValue()==mark)
147:
148:
                      for(int j 0; j<cols; j++)</pre>
149:
                         matrixForMark.addElement(currentRow, ),
150:
                              matrix.getElement(1, j));
151:
                      currentRowit;
152:
153:
154:
             return matrixForMark;
155:
156:
157:
158:
        public int getMaxMark()
159:
160;
             return maxMark;
61:
162:
163:
        public List<String> getAttributes()
164:
165:
             return attributes:
166:
7 A T :
168:
169:
170:
         → dd*precated
771
         * This method is no longer used.
172:
         * diven a list of groups (i.e. sets of feature indices) and the origina.
173:
         * matrix, it would produce a new matrix, with #groups columns, that
174:
         * indicated whether each of the rows in the original matrix 'matched'
175:
         * the group. For example, the nth row in the matrix of [1, 0, 1] would
         * indicate that the nth row of the original matrix 'matched' groups' and
176:
· 77:
         • 3, but not 2. This method was used when group identification was done
178:
         * in 2 stages.
179:
:8u:
        public Marrix getGroupMatrix(List<Set<Integer>> groups, Matrix mx)
181:
187:
             Matrix qMx = new Matrix(rx.numRows(), groups.sizq();;
1月2:
             for int :=0; i<*x.num@ows(); i++)</pre>
194:
185:
                 int
186:
                for(*fetaIntege/> group : groups)
```

```
87:
                     if(matchesGroup,qottup,px,qetK'W 1 '
7 g 4 t
                        gMx.setElement 1, 3, 1...;
                     else gMx.setElement L, 7, 0. 7
193:
o.:
                     2000
* #7:
1935
94:
131
             return gMx;
96:
· ;
138:
       longenepseduitbhis method (referenced above) indicates whether a rew
. 99:No
          ullet natches a particular group. This was done by taking the elements of t^+e^-
200:
          * row occurring in the group, counting the number of 1s, and seeing if
201:
          * inis was greater the root of the size of the group.
202:
203:
        public static boolean matchesGroup(Set<Integer> s, double[] row)
204:
205:
206:
             int matching = 0;
207:
             for(int 1=0; i<row.length; i++)</pre>
208:
                 if(row[1] == 1.0&&s.contains(1)) matching();
209:
210:
211:
             if(new Double(matching)>=Math.floor(Math.sqrt(new Double(s.size()))))
212:
213:
214:
             else return false;
215:
216:
217: }
```

02/20/08 15:00:59

7 I:

--Lor billtarkThad+ 🔭

schemagenerator/ClustererWrapper.java

```
): package sonemagennistin;
 B: import favs..::L.HashSef;
 4: import java.ht.l.linkedbist;
 6: import java.ut 1.Dist;
 c: import rava.util.Ser;
 8: import weka.clusterors.llusterer;
 a: import weki.clusterers.SimpleKMeans;
10: import wexa.core.Attribute;
11: import weka.c.re.FastVectur;
12: import waka.core.Instance;
13: import weka.core.Instances:
14: import weks.core.Matrix:
16: 7**
17: * Provides a less cumbersome way of accessing Weka's clustering
18: • facilities, using the incidence matrix as imput data.
19: * Masthor Jamie
20: *
21: */
22: public class ClustererWrapper {
23:
24:
       Matrix incidence;
35:
       int namClusters;
. 6:
       Consterer plustater = new dampleKMakma();
27:
28:
       public ClustererWrapper(Matrix incldence, int numClusters)
25:
30:
            // We normalise the columns to improve the accuracy
31:
            // of the clustering.
            this.incidence = MatrixUtils.cormaliseColumns(incidence);
32:
            this.ngmClusters ≈ mnmClusters;
33:
34:
35:
               (:SimpleKMeans) of isterar).setNumUlusters :umClusters);
3€:
37:
            ] catch(Exception e)!
38:
               e.printStackTrace();
30:
40:
41:
4.Z:
43:
        * Converts the Matrix to an 'Instances' object, as required
        * by the alustering.
44:
45:
46:
       public void parlaClusterer(List<String> aftrs'
47:
48:
            FastVector fv = new FastVector(attrs.s.ze();
43:
            for Siturg a : antia
달음:
               fv.addElement(new Attribute(s));
51:
             Meed matrix in Instances form.
£2:
53:
            // Each column of the matrix is in instance.
54.1
            Instances ins : new instances "somename", Ft, atims.size(');
··∃:
            for(int i=0; i<indidence:numColumns(); i=40</pre>
               instadd(new Instance(1... noidense.gsf): 1.... 1
56:
7
VB:
            try
., 2
- : :
                Nuasterer, buildClasterer (unst)
0.10
             r catch(Exception e. □
```

```
63:
64:
651
opi
        * Identifies the cluster for each column, and produces a list of sets,
٤:
68:
        * each set representing a group of columns (.a hadev form).
É:
70:
      public fistrSqtrInteger>> getClusters()
71:
72:
            list<Se'<integer>> clusters = new binkedbist<Set<'nteger>>();
73:
           74:
              of tates and id .new FashSet < !nteger>());
75:
70:
           for(int : ; <incidence.humColumns(); i -)</pre>
7 .:
78:
              try
79:
80:
                   int cNum = clusterer.clusterInstance(
8ı:
                      new Irstance(1.0, incidence.getColumn(1))';
82:
                   c.usters.get(cNum).add(1);
83:
              } catch(Fxcention e)[
1)4:
                   e.printStackTrace();
H5:
86:
97:
88:
89:
           return clasions;
90:
91:
92: }
```

60: 61:

return FlustersBySis;

schemagenerator/ClusterUtils.java

```
1: package schemagemenator;
                                                                                           62:
                                                                                          15:
    5: import java.util.Linkedlist;
                                                                                           -54:
                                                                                                 . Returns a vector representing which features occur with h
    4: import sva.util.List:
                                                                                           69:
                                                                                                 ine classer. e.g. A cluster of
    5: import ava.:fil Sel;
                                                                                                 1/1 (0, 1, 3, 5 ) -> (1, 1, 0, 1, 0, 1)
                                                                                                  67:
    7: /**
                                                                                        namFeatures) (
    9: * Provides utilities with regards to clusters (where
                                                                                           68:
                                                                                                      double | gjusterArray = new double[numleatures];
   9: * each cluster represents a group of related features
                                                                                           69:
                                                                                                      for (int : 0; i < numFeatures; i+-)</pre>
   10: * intending to represent one mark of the mark scheme).
                                                                                           70:
                                                                                                         clusterArray[i] = 0.0;
   ": * @author Jamie
                                                                                           71:
   12: *
                                                                                           72:
                                                                                                       for (Integer : : cluster)
   13: */
                                                                                           73:
                                                                                                         clusterArray[i] = 1.0;
   14: public class ClusterUtils {
                                                                                           74:
                                                                                           75:
                                                                                                      return clusterArray;
   16:
          // Return an ordered list of cluster indexes,
                                                                                           76:
   7:
          // such that the first cluster of the list matches
                                                                                           77:
   18:
          // the row to the greatest extent.
                                                                                           78:
   10
                                                                                                      private static double euclideanDistance(double() vArray) | 
          public static List<Integer> getClustersBySimilarity(
                                                                                           79:
   20:
               double() row, List<Set<Integer>> clusters)
                                                                                           80:
                                                                                                              double temp = 0.0;
   21:
                                                                                                              for (int r = 0; : < vArray.length; 1+4)</pre>
                                                                                           81:
   22:
                                                                                           82:
                                                                                                                      temp += Math.pow(vArray[i', 2);
   23:
                                                                                                              temp = Math.sgrt(temp);
               List<Double> cosSims = new LinkedList<Double>();
                                                                                           83:
   ..4:
                                                                                           84:
                                                                                                              return temp;
   25:
               for (int i = 0; ' < clusters.size(); i++) {</pre>
                                                                                           85:
   26:
                  // First turn cluster into incidence array
                                                                                           86:
   27:
                                                                                                  // This gives a vector containing only the feature values
                  double:| clusterArray = getClusterIncidenceVector(clisters.get(i),
                                                                                           87:
row.length):
                                                                                           88:
                                                                                                  // corresponding to features contained by the cluster.
   28:
                                                                                           89:
                                                                                                  // For example, if the cluster is { 0, 1, 4 } and the row
   29:
                                                                                                  // [1, 0, 0, 0, 1], then we obtain the vector [1, 0, 1]
                                                                                           90:
  30:
                                                                                                  public static double() getRowRestrictedToCluster(Sot<Integer> cluster,
               // Now get vector (cosine) similarity
                                                                                           91:
   31:
               // Ignore length of row, since constant 'given the row'.
                                                                                           92:
                                                                                                      double[] row)
   32:
               double cosSiπ. 0.0;
                                                                                           93:
               for (int k = 0; k < clusterArray.length; k++)</pre>
   33:
                                                                                                       double() clusterArray = qetClusterIndidenceVector(cluster,row.length);
                                                                                           94:
   34:
                                                                                                       cosSim += clusterArray(x) * row(x);
                                                                                           95:
   35:
                  cosSim = cosSim / euclideatDistance(clusterArray):
                                                                                           35:
                                                                                                       int count = 0;
   36:
                                                                                           97:
                                                                                                       for (int ) 0; 1 < clusterArray.length; i+·)</pre>
                  cosS:ms.add(cosS.m):
   37:
                                                                                           98:
                                                                                                         if (clusterArray[i] = |.0) {
   38:
                                                                                          99:
                                                                                                              toReturn[count] = row[ii:
   39:
               // Generate list clustersBvSim
                                                                                          100:
                                                                                                              count+ ::
   40:
                                                                                          101:
   41:
                                                                                          0.2:
              List<Integor> clustersBySim = new LinkedList<Integer>();
                                                                                                      return toReturn;
   42:
                                                                                          103:
   43:
               double lastMax = 10000;
                                                                                          104:
                                                                                                  public static int getClusterVectorSize(double[] clusterArray) {
   44:
               while (clustersBySim.size() < cosSims.size()) {</pre>
                                                                                          105:
   45:
                  int ourrentMaxCluster = -1;
                                                                                          106:
                                                                                                       int oneCount 0;
   46:
                  double ourrentMax = -(.0;
                                                                                          C7:
                                                                                                       for (int i = 0; i < clusterArray.length; '++)</pre>
   47:
                  for (int 1 = 0; 1 < cosSims.size(); 1++) (</pre>
                                                                                                         if (clasterArray[1] == 1.0)cneCount(+;
                                                                                          80
   48:
                                                                                          109:
                       if .cosSims.get(1) >= currentMax && cosSims.get(1) < lastMax</pre>
                                                                                                       return oneCount;
   49:
                           &% !clustersBySim.contains(1))
                                                                                          110:
  50:
                                                                                          111: ]
  50:
                          currentMax = cosSins.get(1);
  52:
                         currentMaxCluster = 1;
  53:
  54:
  55:
                  clustersBySim.aco(gurientMaxCluster);
  56:
                  lastMax - CurrentMax:
  57:
  56:
  59:
              // System.out.println("Cos sims: "-cosSims);
```

15:10:38

schemagenerator/CumulativeClassifier.java

3 ... y

```
63:
 ': package schemagenerator;
                                                                                                      Matrix instrussfrrendrim\mathbf{x} = \text{Matr} \mathbf{x}' tile
                                                                                                               .derInbidenceMatrix | hetance:Reader
 : import have.i..Edin=Write:;
                                                                                          65:
                                                                                                                       .getMatrixFrom ha anima instances .:
                                                                                          66:
 4: import ravalud h.bist;
                                                                                          67:
 h: import have, in 1.8et;
                                                                                                      maxMark: InstancesReader.getMuxMark(instances);
                                                                                          5e:
                                                                                                      if we haven't set the total number of clusters,
                                                                                          69:
                                                                                                      if (totalC'usters == 0)
 /: import weka.classifiers.Classifier:
 A: import weka, plassifiers.payes. NatveBayes;
                                                                                          70:
                                                                                                              totalClusters = maxMark:
 9; import weka.dore.Attribute;
                                                                                          72:
                                                                                                      List<String> attributes = InstancesReader.getAttributes(instances);
13: import wekaloure.FashVectur;
11: import weka.core.Instance;
                                                                                          735
12: import weka.core.Instances;
                                                                                          74:
                                                                                                      if (incomfile = null)
                                                                                          15:
                                                                                                              viosters Group-dentifie: .identityGroupsJsingClustering(
13: import weks.core.Mairix;
14:
                                                                                          76:
                                                                                                                               iestureStrengthMx, attributes, maxMark):
15. /**
                                                                                          17:
                                                                                                      // else clusters SchemaUtils.readMarkSchemeFile(inputFile,
                                                                                          78:
16: * A classifier which takes into account the cumulative nature of a
                                                                                                      / attr_butes):
1/: * classification, that is, the classifications x and y if 2 independent
                                                                                          79:
                                                                                                      else
in: * properties X and Y are satisfied correspondingly should result in a
                                                                                                              clusters - Schemadtils.readMarkListFile(inputFile, attributes.
                                                                                          80:
19: * classification of xiv if both X and Y are satisfied.
                                                                                          g1.
                                                                                                                       totalClusters):
                                                                                          82:
201 =
21: * @author Jamie
                                                                                          ម::
                                                                                                      class flors = new Classifier!fotalClusters):
22: *
                                                                                          84:
                                                                                                      DistancesArray = new Instances(total/lusters):
                                                                                          85:
24: public class Cumular.veClassifier extends Classifier (
                                                                                          86.
                                                                                                      // STEP 0:
                                                                                                      // Output mark scheme if necessary.
      // The classifiers for each mark.
                                                                                          87:
      Classifier! | classifiers:
                                                                                          58:
                                                                                                      if (hathutFile != null) !
      // The data set relating to each mark.
                                                                                          89:
                                                                                                              System.out.println("Outputting mark scheme to: "
                                                                                          90:
      Instances[] instancesArray;
                                                                                                                       + outputFile);
                                                                                          91:
      // The features relating to each mark in the mark scheme.
                                                                                                              FrintWriter pw = new PrintWriter(outputFile);
       List<Set<Integer>> clusters:
                                                                                          92:
                                                                                                              pw.println(attr(buies):
      int totalClusters = 0:
                                                                                          93:
                                                                                                              for (Set<Integer> c) uster : clusters)
      // Note that the maximum maik is not necessarily
                                                                                          94:
12:
                                                                                                                      pw.println(cluster);
      // the same as the number of clusters (i.c. if the mark
                                                                                          95:
                                                                                                              pw.ciose():
      // is limited to a maximum below the rotal feasible marks).
                                                                                          96:
                                                                                         97:
       int maxMark ≈ 0:
                                                                                                      // STEP 1:
                                                                                          98:
A6:Outputs the mark scheme to a .markscheme file.
                                                                                         99:
                                                                                                      // Initiate the classifiers, and start a fresh instances object
      String outputfile : null;
                                                                                         100:
                                                                                                      // for each of the classifiers, each with the correct attribute
       // Optionally inputs a manually constructed .marklist
       // file, which overgides the clustering and specifies
                                                                                         161:
                                                                                                      // information.
       // which mark each feature corresponds to.
                                                                                        102:
41:
       String inputFile = null;
                                                                                         1331
                                                                                                      for int i = 0; i < totalClusters; i+-) {</pre>
                                                                                         104:
.2:
                                                                                                              oldesifiers[i] = new NaiveBayes();
       public CurulativeClassifier() {
                                                                                         135:
43:
                                                                                                              // classifiers(i) | new Decisi mTable();
                                                                                         106:
44:
                                                                                        107:
43:
       public CommulativeClassifier(String inputFige, String outputFile) {
46:
                                                                                         108:
                                                                                                      for int = = 0; = < maxMark; --+)</pre>
47:
            this.imput: le = irputF.le;
                                                                                         109:
48:
            this.butputFile = outputFile;
                                                                                         110:
                                                                                                         int sizeOfCluster = ClusterUtils.getClusterVeitorSize(ClusterUtils)
                                                                                         111:
. c
                                                                                                              .getClusterIndiaenceVector(clusters.get(i), attributes
                                                                                         112:
E::
                                                                                         -13:
                                                                                                         PastVector iv = new FastVector(sizeOfCluster + 1);
51:
       public CumulativeClassifier(String inputF.le, String outputFile,
                                                                                         1.4:
57:
            int totalClushers'
                                                                                                         for (int j = 0; j < sizeOfCluster; \( \pm \);</pre>
53:
                                                                                         115:
                                                                                                              fv.adoFlement(new Attribute("attr" = - -));
54:
                                                                                         116:
            this(inputFile, .orpatFile);
55:
            this t falliusters = total Clusters;
                                                                                        117:
                                                                                                         / The classification is a NOMINAL. Either 0 or 1.
                                                                                         1.8:
56:
                                                                                                         -/ therefore need to make oroup.
                                                                                        119:
57:
                                                                                                         FastVector binary = new FastVector(2);
ამ:
       public void PuljdC.assifier(Instances instances: throws Exception i
                                                                                         120:
                                                                                                         pilary.adoElement."0";
                                                                                         121:
                                                                                                         Libary.aduflement("1");
59:
411
                The festure strength matrix is an axa matrix (where there are a
                                                                                         122:
                                                                                                         fv.additlement (new Artripute "score", cimary :
71:
            \sim (eafures) such that M(1,7) indicates the number of those seasones.
                                                                                         .3:
                                                                                                         instancesArray[if = new instances "iSet" = i, iv. ng.ng.
· 2:
                 ordin arcear i Wether in the dataset.
                                                                                         124:
                                                                                                         instancesAcray(i)
```

schemagenerator/CumulativeClassifier.java

```
125:
                      .setClassIndex(instancesArray[i].numAtrributes() 1);
106:
127:
128:
- 29:
              // Now populate the instances object for each of the classifiers.
136:
131:
              Matrix OMx = InstancesReader.getMatrixPromInstances(instances);
1.12:
:53:
              for (int 1 0: 1 < oMx.numsews(); 101; {</pre>
134:
135:
                 double[] row = oMx.getRow());
136:
137 -
                 // Need to find clusters it's most similar to
138:
                 List<Thteger> clustersOrdered ClusterUtils
139:
                       .getClustersBySimilarity(row, clusters);
140:
                 int classification = InstancesReader
41:
                       .getClassification(instances, i);
142:
143:
                 // We take the first 'classification' clusters from ordered
144-
                 // clusters, and train the networks for these clusters.
145:
                 for (int ) = 0; j < clusters.size(); j++) (</pre>
146:
                      // Train cluster
747:
148:
                      // But first, need to construct instance!
149:
                      // To do this, we simply map incidence vector for cluster
150:
                      // on the row.
151:
                      double[] clusterArray =
152:
                         ClusterUtils.getRowRostrictedTcCluster(
153:
                              clusters.get(c.ustersOrdered.get(t)), row);
154:
                      Instance instance = new Instance(clusterArray.length + 1);
155:
                      instance.setDataset(instancesArray[clustersOrdered.get(7)]);
156:
                      for (int k = 0; k < clusterArray.length; k++)</pre>
157:
                         instance.setValue(k, clisterArray[k]);
158:
                      if (i < classification) {
159:
                         instance.setValue(clusterArray.length, "1");
160:
                      l else f
161:
                         instance.setValue(clusterArray.length, "0");
162:
163:
164:
                      instancesArray[clustersOrdered.get())].add(instance);
165:
166:
67:
68:
169:
170:
              // STEP 3:
771:
              // Finally, build the classifiers using the instances objects.
1/2:
173:
              try [
174:
                 for (int i = 0; _ < totalClusters; i++)</pre>
175:
                      classifiers(i).billdClassifier(instancesArray '');
1/6:
              } catch (Exception e) {
1//:
                 e.printStackTrace();
178:
179:
180:
              // For the user to see, print the mark scheme
181:
              SchemaUtils.print(clusters, actrioutes);
182:
193:
164:
        public double classifyInstance(Instance instance)
135:
                      throws gava. (ang.Extention {
186:
              // For each classifier, we need to:
```

```
// 1. Restrict the instance only to the attributes it contains.
 67:
             .. 2a. Do some higgery pockery to make the instance suitable for
188:
182:
                c'assification.
196:
             // 26. Classify this restricted instance.
191:
             33. Add together all classifications.
              - 3b. Cap the mark ar maxMark.
197:
93:
             double total = 0.0;
194
145
196
             for (int 1 = 0; 1 < classifiers.length; i-+) ;</pre>
197:
                     double [ restrictedInstance = Clusterttils
198:
                        .getRowRestractedToClustor(clusters.get(i), instance
139
                             .toDoubleArray());
200:
                     double [] withArbitraryClassification = appendClassification(
                        restrictedinstance, 2.0);
201:
202:
                     Instance instanceB = new Instance(*.0,
                             withArbitraryClassification);
203:
                     instanceR.setDataset(instancesArray[:]);
204:
205:
                        double d = classifiers[1].classifyInstance(instanceB);
206:
207:
                        total + d;
                     ] catch (Exception e) {
208:
                        throw new Error(e);
209:
210:
211;
212:
             if (total > maxMark)total = maxMark;
213:
214:
215:
             return total;
216:
        ì
217:
        private double[] appendClassification(double[] cArray, double toAppend) {
218:
219:
             double[] cArrayNew = new double[cArray.length + 1];
220:
             for (int i = 0; \pm < cArray.length; i++)
221:
                     cArrayNew[]] cArray[];
             cArrayNew(cArray.length) LoAppend;
222:
223
             return cArrayNew;
234:
225:
226: 1
```

55:

else

schemagenerator/GroupIdentifier.java

```
: package shiosasemerathr;
                                                                                            , ´:
                                                                                                                 Perkhat Priedorsky new bulls in new Harmaink Tet Stiggters, A:
    3: import mava.uni .diaanden;
                                                                                                                 boolean dudatAppend = true:
    4: import dava.html.limkedlest;
                                                                                            62:
    5: import nava..til.List;
                                                                                            . 5:
                                                                                                                 for detkinhegors an un i abrupon'
    6: import hava.uni .San;
                                                                                            F .
                                                                                            oê:
                                                                                                                    SetkInteger> Sarundwords = new HashSetkInteger> ':
    8: import weka.core.Matrix;
                                                                                            (6:
                                                                                                                    Set<Integer> weakBonds = new HashSet<Integer>();
                                                                                            67:
   9:
  10: /**
                                                                                            681
                                                                                                                    for Integer 7 : group,
  1): * This class supplies different methods to includy
                                                                                            6,9:
                                                                                                                        if(j! 1)
   12: * groups of features based on the features' co-occurence.
                                                                                            70:
                                                                                                                            if featureStrengtnMx.getEichenf(_,-)>MINIMUM POWD(s
  13: * The intended output of such methods is a list of sers,
                                                                                         orungBandriada 1:
   14: * where each set represents a group of features (in index
                                                                                            71:
                                                                                                                                  else weakBonds.add(1);
   15: * form).
                                                                                            , Z:
                                                                                            73:
  16: * Wauthor Jamie
   · */
                                                                                            74:
                                                                                                                    // System.out.println("Strong bonds: ".strongBonds);
                                                                                            75:
  18:
                                                                                                                    // System.out.println("Weak bonds: "+weakBonds);
                                                                                            76:
   19: public class Groupidentitiet
                                                                                            77:
   2:: :
                                                                                                                    // If weakBonds is empty, then the element becomes part of
   31:
                                                                                          that set.
          * This method uses standard (simple K-means) clustering.
                                                                                            78:
                                                                                                                    if(weakBonds.isEmpty())
   27:
                                                                                            79:
   23:
          public static fast<Set<Integer>> identifyGroupsUsingClustering(Matrix Featu
                                                                                                                         group.add(1);
reScrength.Mx, list<String> features, int numClusters)
                                                                                            81:
                                                                                                                         didntAppend : false;
                                                                                            82:
   25:
                                                                                                                     else if(strongPonds.isEmpty())
                                                                                            83:
               rear graStrengthMx = MatrixUtils.norma isploMax(featureStrengthMx);
   26:
                                                                                                                         // It strongBonds is empty, then we need to create a L
  27:
               ClustererWrapper cw = new ClustererWrapper(teatureStrengthMx, numClust
                                                                                            84:
ers):
                                                                                         ew set
               ow.busidClasterer(features);
                                                                                            85:
                                                                                                                         // containing just that feature. HOWEVER, this only oc
  25:
               return ow.getClusters();
   29:
                                                                                         CHI'S
                                                                                            86:
   30:
                                                                                                                         // if the feature was not added to any set.
                                                                                            87:
  31:
                                                                                            45:
                                                                                                                    ; else {
   32
          // Put in mind matrix is normalised.
                                                                                            89-
          // Anything under value is considered a 'weak boni'
                                                                                                                         // Both strongBonds and weakBonds is non-empty.
   33:
                                                                                            90:
   34 :
          // Anvibing above is consider 'strong'.
                                                                                                                         // We simply make a new group with the strongBonds
   35:
          final static double MINIMUM BOND 0.4;
                                                                                            91:
                                                                                                                         // and the element to add. The original group is
   30
                                                                                            92:
                                                                                                                         // unaffected.
                                                                                            93:
   37:
                                                                                                                         strongBonus.add(,);
          * @deprecated
                                                                                            94:
   36:
                                                                                                                         newGroups.add(strongBonds);
           * Produces the groups by identitying cliques (where feature
                                                                                            95 -
   39:
                                                                                                                         didntAppend = false;
           * nodes have an edge between them if their connection is
                                                                                            96:
  4G:
                                                                                            97:
   41:
           is sufficiently strong) before wording them into the desired
                                                                                                                 ŀ
                                                                                            98:
   42:
           * number of droves.
   43:
                                                                                            99:
                                                                                                                 if (as dntAppend)
          public static Set<Set<!r-eder>> identifyGroupsUsingCliqueMenging(Matrix Fea
                                                                                            100:
   41:
two-direngthMx)
                                                                                            01:
                                                                                                                    SetkInteger> s = new HashSatkIntager>( ;
                                                                                            102:
  45:
                                                                                                                    s.add(i):
                                                                                           103:
               fearureStrengthMx = MarrixUtils.thrmaliseToMax(featureStrengthMx);
   46:
                                                                                                                    newGroups.add(s);
                                                                                            04:
                                                                                            195:
   47:
               SetkBotkInteger>> groups = new FashSotkSetkIntegor>> 1;
                                                                                                                 croups.aadAll(newGroups):
   45:
                                                                                           106:
   49:
               for(int 1.0; 1<featureStiengtnMx.numkows(); +++)</pre>
                                                                                            101:
                                                                                            108:
   50:
                                                                                                        return aroups;
                  System.cut.routtln?"Processing row "+1';
                                                                                            150:
  51:
  52:
                  if (groups.lsEmpty.c)
                                                                                            . . .
                                                                                           111:
   53:
   54:
                       // Simple start case
                                                                                           1 24
                       workStredmi> a = new Mast.Setkinteber> ;
   55:
                                                                                                   public static void main(String[] args)
                                                                                           ं भ:
   56:
                       stand our
   57:
                       uroups.com.s.r
                                                                                                            Table the above aroup dientification
                                                                                           120;
```

177:

rxi.setElemost(1,4, 0.0);

schemagenerator/GroupIdentifier.java

```
117:
                                                                                                            txi.serflerent 1,5, \.0 ;
                ListeString> features = new Linkeolist<String>();
                                                                                               5
  ...5:
                fear .pas.add "A"\:
                                                                                               179:
                                                                                                            mx^2.set filement(2,2, 1.3);
   ∵∌:
                restures.adu."B".;
                                                                                               . .
                                                                                                            rxi.serElement 2,3, 100 ;
  · · · · ·
                teatures.add/"C"):
                                                                                               131:
                                                                                                           ux3.setElement(2,4, 0.0.;
                                                                                               a,
                tear neal said. "D" .:
                                                                                                           hx3.setElement(2,5, 3.0);
  122:
                features.add("E");
                                                                                                            mxs.serileennatus,3, .....
                                                                                               183:
  723:
                icatures.ado "F" ;
                                                                                               184:
                                                                                                            mxs.setElement(3,4, 0.0);
  724:
                                                                                                            mxp.serElement.p.,5, 0.0';
                Teatures.add("G");
                                                                                               655
  25:
                featires.adi "#";
                                                                                               146:
                                                                                                           mx+.setflement(4,4,-1.5);
  126:
                                                                                                            tx3.setflemest(4,5, 1.0);
                                                                                               5.
  27:
                Matrix mx = new Matrix(4,4);
                                                                                               188:
                                                                                                            mx3.sctflement(5,5, 1.0);
  128:
                mx.setSlement(0,0, 100);
                                                                                               169:
                                                                                                           rugrarMx(mx3);
  129:
                mx.setFlement(0,1, 0.0);
                                                                                               .00:
  130:
                                                                                                            groups = Group!dentifier.identifyGroupsUsingCliqueMerging(mx3);
                mx.seti'emon'(0,2, 1.0);
                                                                                               191:
  131:
                                                                                                           groups - Merger.mergeGroups(proups, mx3, 2);
                mx.setElement(0,J, 1.0);
                                                                                               192:
  132:
                mx.setElemant(1,1, 1.0);
                                                                                               193:
                                                                                                           SchemaUtils.prunt(groups, features);
                                                                                               34:
  133:
                mx.setFlemont(1,2, 1.0);
  34:
                mux.setElement(1,3, 0.0);
                                                                                               195:
                                                                                                            System.out.println();
  135:
                πx.setElement(2,7, 1.0);
                                                                                               196:
                                                                                                            ArifReader or = new ArifReador("./files/13bir-0.arff");
  136:
                mx.setElement(2,J, 1.0);
                                                                                               197:
  137:
                cx.setElement(3,3, 1.0);
                                                                                               198:
                                                                                                            Matrix mx4 = MatrixUtils.getIncidenceMatrix(ar.getMatrix());
  138:
                mirrorMx(mx);
                                                                                               99:
                                                                                                            bx4 = MatrixUtils.normaliseToMax(mx4);
  139:
                                                                                               200:
  140:
                                                                                                            groups = GroupIdentifier.identifyGroups(mx4);
                Set<Sct<fnteger>> groups = GroupIdemrifier.idemrifyGroupsUsingCliqueMe
                                                                                               201: //
raina(mx):
                                                                                               202: //
                                                                                                            SchemaPrinter.print(groups, ar.getAttributes());
  41:
                Schemattrils.print(groups, features);
                                                                                               203: //
                                                                                                            System.out.println("Merging");
  142:
                                                                                                            groups = Morger.mergeGroups(groups, mx4, ar.getMaxMark());
                                                                                               204: //
  143:
                System.out.println("");
                                                                                               205: //
                                                                                                            SchemaFrinter.print(groups, ar.getAftributes());
  144:
                                                                                               206:
  145:
                Matrix mx2 = new Matrix(5,5);
                                                                                               207:
                                                                                                            List<Set<Integer>> clusters = identifyGroupsUsingClustering(mx4, ar.qe
  146:
                mx2.setElemont(0,0,1.0);
                                                                                             tAttributes(), ar.getMaxMark(');
  47:
                Tx2.setElement(0,1, 1.0);
                                                                                                            SchemaUtils.print(clusters, ar.getAttributes());
                                                                                               208:
  148:
                                                                                               109:
                mx2.setElement(0,2, 1.0);
  149:
                                                                                               210:
                mx2.setFlement(0,3,0.0);
  155:
                mx2.setFlerent(0.4, 0.0);
                                                                                               711:
  151:
                mx2.setElement(1,1, 1.0);
                                                                                               212:
                                                                                                        * Using by the demo to copy a triallular matrix into
  152:
                \texttt{Tx2.setElement(1,2, 1.0)};
                                                                                               213:
                                                                                                        * its diagonal reflection so that the resulting matrix
  153:
                mx2.setElement(1,3, 0.0);
                                                                                               2~4:
                                                                                                        * is symmetric.
  54:
                mx2.setElement(1,4, 0.0);
                                                                                               215:
  155:
                                                                                                       private static void mirrorMx(Matrix Mx)
                mx2.setFlement(2,2, 1.0);
                                                                                               7.6.
  156:
                mx2.setFlement(2,3, 1.0);
                                                                                               217:
  157:
                                                                                                            for(int r = 0; r < mx.numCclamns(); r = +)</pre>
                m \times 2.setFlement(2.4, 1.3);
                                                                                               218:
  158:
                mx2.setElement(3,3, 1.0);
                                                                                               719;
                                                                                                               for(int c= ; c<*; ---
  155:
                                                                                               220:
               nx2.setslemont(3,4, 1.0);
                                                                                                                    mx.setElement(r,c,mx.getElement(c,r));
  160:
                mx3.setFlement(4,4, 1.0);
                                                                                               221:
  161:
               ParcorMx (mx2);
                                                                                               /22:
  162:
                                                                                               223: 3
  763:
               groups = Group?dentifier.idontifyGroupsUsingCliqueMerqing(mx2);
  64:
                Schetaitils.print(groups, features);
  165:
                System.out.println("");
  6.:
  167:
                Matrix mx3 new Matrix(6,6):
               mx3.setElement(0,0, 1.0);
  ်စ်စ:
  169:
               mx3.setEiemens(0,1, 1.0);
  173:
               mx3.setElement(0,2, 1.0);
  17 :
               mxi.setElement(0,3, 1.0,;
  1/2:
               mx3.setElement(0,4,0.0);
  · •:
               mxilsenE enema( ,E, 0.0);
  4:
               mx 1.ser2'esent(1,1, 1.0);
  ^ ~ ÷ :
               "x .set Lemont(',2, '.3);
                ixise vieromt(1,3, 1.1);
```

schemagenerator/InstancesReader.java

60: 62:

6-:

27.2

67:

on: .

```
7 A . S
```

```
: package schemagenerator;
 o: import nava.ltil.tj.nkcd.tet;
 4: import hava.util.List;
 e: import weka.core.Autribute;
 7: import wexalders.Instance;
 b: import weks.core.Instances;
 9: import weks.core.Marrix;
11:
11: ***
12: * Whereas ArifRoader oprains the representative matrix and other
's: * data from an Arff file, this provides similar extraction from
'4: * an Instances object.
18: * @author Jamie
16: *
1/: */
18: public class InstancesReader {
       public static Matrix getMatrixFromInstances(Instances Instances)
2:
            Matrix mx = new Matrix(instances.numInstances(),
21:
22:
                    instances.numAtt::putes(,-',:
23:
24:
            for(int row = 0; row < instances.numinstances.); row++)</pre>
            f Instance ins = instances.instance(row);
25:
20:
               double[] insArray = ins.!oDoubleArray();
27:
               for(int col = 0; col < instances.numA(tributes()-1; col+-)</pre>
28:
29:
                     mx.setElement(row, col, ins.value(col,);
30:
31:
32:
            return mx;
33:
34:
35:
36:
       public static int getMaxMark(Instances instances)
3 / 2
38:
            double maxMark = 0.0;
39:
            for(int row = 0: row < instances.numlastances.); row:-)</pre>
4.:
               Instance ins = instances.instance(row);
41:
               double() insArray = uns.reDoub(eArray();
42:
               double classification insArray(instances.numAttributes() - 1;
               maxMark = (classification>raxMark) ? classification : rexMark;
45:
16;
47:
            return (new Double, muxMatk (.antValue );
48:
49:
50:
       public static [ist<String> getAltributes(Instances instances)
1<sub>2</sub>1 :
52:
            listkString> attributes = new linkedutstkString>');
53:
            // Don't want the last altiibute (which is 'score')
            for(int i=0; ikinstances.turAttricutes,'-'; i--)
- 4
55:
56:
               Altribate att = instances.attribute(1);
5 1:
               autrioures.addyantiname '';
58:
59:
            return attributes;
. 0
б°:
       public static into set Table : hunting Thanking estances, into index
```

62:

The Charles

schemagenerator/MarkSchemeLister.java

```
1: package programmentation:
                                                                                         63:
                                                                                                        Pur in map, with all instialized to empty (sale (i.e. no aroug))
                                                                                         - 14
                                                                                          a5:
                                                                                                        for Sig to attn : attn.outes
 or import (avail logoforodResdet)
                                                                                                             groupMappings.put.attr, new TreeSonkIncolf(2). );
                                                                                          66:
 4: import flavilendeledeader:
 or import rand. LinguistreadReader;
                                                                                          ř^:
                                                                                                     else
                                                                                                        . Read Darkijst ille
 of import was assert TWinter;
                                                                                          ≎0:
                                                                                                        gruccMappinus = BenemaUtils.getGrucpMappings.irp.dis.e.;
 69:
                                                                                          70:
 8: import available, ist:
                                                                                          7::
 A: import | Ava. . . Nati:
11: import java.ut.1.Set;
                                                                                         72:
                                                                                                      / Provide a command line interface to modify features' moups.
": import javalut,...TreeMap;
                                                                                          73:
                                                                                          74%
                                                                                                     System.out.orontlm("Commands:");
12: import ava. :::1.TreeSet;
                                                                                          75:
                                                                                                     System.bit.println("\tASSIGN attr groups";
                                                                                                     {\tt System.out.println("\tCONTAINS word groups"):}
4: /**
                                                                                          76:
                                                                                         1.15
                                                                                                     System.out.println("\tQUIT");
15: * Used to convert a markscheme file into a marklist file.
                                                                                                     78:
16: * A marklist file is a more convenient form for manual
17: * editing, and has lines of the form:
                                                                                          79:
                                                                                                     PufferedReader br = new BufferedReader(
18: * "feature [list of groups-if-appears-in]"
                                                                                          80:
19: * If a file main.marklist is placed in a test directory,
                                                                                          81:
                                                                                                             new InputStreamReader(System.in));
20: * this will override the mark groups.
                                                                                          82:
                                                                                                     boolean guit = false;
                                                                                          83:
21: * @author Jamie
                                                                                          84:
                                                                                                     while(!ourt)
22: *
                                                                                          85:
23: */
                                                                                                        System.out.print("> ");
                                                                                          86:
24: public class MarkSchemeLister (
                                                                                          87:
                                                                                                        String line = br.readLine();
                                                                                          88:
26:
       static Man<String,Set<Integer>> groupMappings =
                                                                                                        String[] lineParts = line.split(" ");
27:
            new TreeMap<String,Set<Intoger>>();
                                                                                          89:
                                                                                                        if(lineParts[0].equals("ASSIGN"))
28:
                                                                                          90:
29:
       public static void main(String() args) throws Exception {
                                                                                          91:
                                                                                          93:
                                                                                                             String[] groups = lineParts[2].split(",");
30:
            if(args.length =0)
                                                                                          93:
                                                                                                             for(int j=0; j<groups.length; j++)</pre>
31:
                                                                                          94:
                                                                                                                groupMappings.get(lineParts|1 ).add(
32:
              System.out.println("USAGE:");
                                                                                                                     new lateger (groups[]]);
                                                                                          95:
33:
              System.out.print("\tMarkSchemeLister in.markscheme out.marklist");
                                                                                                             output(outputFile);
34:
              System.out.println("\t\t or");
                                                                                          96:
35:
                                                                                          97:
              System.out.println("\tMarkSchemeLister in.marklist out.marklist");
                                                                                                        if():neParts[0].equals("CONTAINS"))
36:
              System.exit(0);
                                                                                          98:
37 •
                                                                                          99:
                                                                                                             String[] groups = lineParts[2].split(",");
38:
                                                                                         100:
                                                                                         101:
                                                                                                             for(String attr : groupMappings.keySet())
39.
            $tring inputSile = args[];
                                                                                                                if(attr.contains(lineParts[1 )){
                                                                                        102:
40:
            String cutputFile = args[1];
                                                                                                                     for(int j=J; j<groups.lengin; j++)</pre>
                                                                                         103:
41:
                                                                                                                       groupMappings.get(attr,.add(new Integeriatoups[5]));
42:
             / Read in
                                                                                         134:
                                                                                         1.5:
                                                                                                             output (outputFile);
43:
            if input: o.commains:"markscheme";)
                                                                                         196:
44:
                                                                                         137:
                                                                                                        if(lineParts[0],equals("QUIT")) dust = true;
45:
               BufferedReader or = new BufferedPeader(new rolePeader(orgutFole));
                                                                                         138:
46:
                                                                                         109:
                                                                                                     pr.close ';
4 \cdot z
               Bissing acting - priceedDing();
               int maxMark 1;
                                                                                         1101
48:
                                                                                        ٠...
49:
               while (br. ready (1))
                                                                                          12:
                                                                                                     public static void putput String thleName throws Rx Megling.
აე:
                    #axYark-+:
                                                                                        · · ¿:
51:
                    or.readline();
                                                                                        ....
                                                                                                             PrintWriter bw = new PrintWriter(fileName ;
5.7 *
                                                                                         115:
                                                                                                              for(String attr : groupMappings.keySet())
53.
               Brichesco;);
                                                                                         116:
5.4 :
               System.out.printin("NOTE: There are "+paxMark+" groups.");
55:
                                                                                         117:
                                                                                                                     bw.print(attr+"\t");
                                                                                         . ។ មៈ
                                                                                                                     pw.print(groupMappings.geflatt* );
56.
               · · Now put attributes in list.
                                                                                        1-9:
                                                                                                                     pw.printlh();
57:
               attrs = attrs.substring(1,attrs.length()-1);
                                                                                         126:
58:
               list<String> attributes = new LinkedList<String>();
                                                                                         101:
50:
               String Olivez = antra.spirt(", ");
                                                                                                             pw.diose();
               for(int =0; ikattiaz.length; i++)attributes.applattrs2.ii);
60:
                                                                                         122:
                                                                                         123: 3
67:
```

schemagenerator/MarkSchemeModifier.java

```
1: package schemagemenator;
    9: import fav...i.saferedPeadat;
    4: import rava.io.hiloReader:
    5: import java.id.anpinifrananReauer;
    6: import java.gt..runtWourer;
    7: import hava.btir.HashSet;
    8: import java.utl..LinkecList;
    3: import -ava.uti.list;
   10: import -ava.util.Set;
  11:
   12: /**
   ::: * Adeprecated
   14: * dauthor Jamie
   16: */
   17: public class MarkSchemeModifier [
               public static void main(String) args) throws Exception (
   19:
                       if (args.length= 0)
   20:
                               System.out.pr http://USAGE:");
   2.
                               System.out.println("\tMarkSchemeModifier in.markscheme
   22:
out.markscheme");
   23:
                               System.exit(0);
   74:
   25:
   26:
                       String importain = args[3];
                       String outputFile ards[1];
   27:
   28.
                       BufferedReader or new BufferedReader(new FileReader(.mputFil
   29:
e),;
   20
                       String aits = br.readLine();
                       int maxMark = 0;
   31:
   32:
                       while (br. realy()) !
   33:
                               maxMark++;
                               br.readLine();
   34:
   30.
                       br.close():
   36:
                       System.out.println("NOTE: There are "-maxMark+" groups.");
   37:
   36:
                       // Now put attributes in list.
   33:
                       attis = attrs.substring(),attrs.length(,-1);
                        _istkBtring> sttrigutes = new linkhdlistkString> ';
   41:
                       String[ attrs/ = aftrs.splif(", ");
   -2:
   43:
                       for(int i=0; ikattrs2.length; i++)attributes.add(attrs2[ii];
   44:
                       System.put.printle("For each attribute enter a group from 1 to
   d_2^2 = \frac{1}{2}
 "-naxMark." to place it in.");
                       Fyster.cor.probable ("For multiple groups, separate by spaces."
   46:
                       Sys(em.out.println("-----"
   47:
   48:
   49:
   50:
                       // Now get new groups
                       SetkInteger> ] newGroups - new Set HaxMark!;
   5 1
                       for(int n=0; nemaxMark; n++)newGroups(n) = new HasnSerkIntegs;
   >2∶
>():
   53:
   ~ 4 :
                       or = new Butfereakeader (new imputStreamSeero: Systi.in )
   55:
   10
                       for int (); rearizable residue (); in-
```

```
58:
                              System. hat. printing acts outes .get( ):
                              82 min., 15.82.55 "> " ;
  59:
  A .
                              Etrans provogelectuar = br.read[inc]);
  61:
                              if (!aronoS-lection.equals: ""):
  633
                                      String[] divers = proupSelection.sp(r(" ");
  63:
  64:
                                       for(int j=0; j<qramps.length; t++)newGroups[ne</pre>
 Integer (groups[]]) - 1 (.340 (1);
  65:
  66:
                      System.out.println("-------
  69:
   16:
                      System .out.println("Done, writing to file.");
  71:
  12:
                      PrintWriter pw : new PrintWriter (outputFile);
  73:
                      ow.println(autr.butes);
  74:
                      System.out.println(altributes);
  75:
                      for(int i 0; i<maxMark; i++){ pw.println(newGroups[i]); Systems</pre>
.out.printin(newGroups(i)); }
  76:
                      ow.close();
  77:
  78: 1
```

schemagenerator/MatrixUtils.java

```
1: package schemagemenator;
 3: import waka.core.instance;
 4: import waka.sare.instances;
 5: import see., pire.Manauk)
 8: * Utilities relating to matrix manipulation.
 i: * Mauthor Jamie
16: **
The public class Matrix It is [
72:
       public static Matrix nortaliseToMax(Matrix nx)
13:
14:
            // We take the log of all the values.
15:
            // This ensures very high values do not scale
16:
            // everything else down.
11:
            Then we scale everything so that the maximum is ".
18:
19:
            Matrix copy = (Matrix) mx.clone();
20:
21:
            // Find max
22:
            double currentMax = 0.0;
23:
            for(int i=0; icmx.numRows(); i++)
24:
            { for(int ) 0; 7<mx.numRows(); 3++)
25:
26:
                    // Srt element to the log of it.
27:
                    // We plus I since log(0) is undefined.
28:
                    // Therefore U -> 0 as wanted.
29:
                    copy.setElement(1,1,Math.log(copy.getMlement(1,1).1.0));
30:
                    if(copy,getEloment(1,j)>currenLMax)
31:
                       currentMax copy.getElement(1,j);
32:
33:
34:
35:
            7: Divide everything by max
36:
            .. Diagonals are set to 1.0, such that a feature will always match
37:
            // with itself.
38:
            for (int 1=0; _<mx.numRows(); 1++)</pre>
39:
               for(int = 0; f<mx.numRows(); j++)</pre>
40:
                    if(i1-j)copy.setElement(i, j,
4`:
                       orpy.getEloment(1,3) / durientMax';
42:
                    else copy.setElement(1,3, 1.0);
43:
44:
            return copy;
45:
46:
4/:
       // Obtains [AAT] A
46:
       public static Matrix getIndiaendeMatrix(Matrix mx) {
44:
            return mx.transpose().multiply(mx);
56:
51:
       public static Matrix horrallseColumns Matrix fx
52:
⊃3:
54:
            Matrix mx2     new Matrix(mx.mumRows(), mx.mumColumns());
55:
            // Normalises columns
Se:
            for(int l=0; i<mx.tumColu.ns(); i++)</pre>
57:
               double squaredSum (0.0;
ō₩:
5 1:
               for(int plant pknx.nunR ws/ ; jak
10112
                    squaredSum+= Muln.pow(mx.perE ement(), i), 2.5);
               squaredSum - Math.sqrt(aquaredFum);
- 1:
· 2:
```

schemagenerator/Merger.java

~ /- :

たっこ

opi

હતે∶

70:

7.5 ±

72:

73:

74:

75:

76:

78:

19:

80:

81:

82:

83:

84:

85:

86:

87:

សស:

89:

90:

91:

92:

14:

95:

36:

97:

98: 99:

100:]

```
1: package schemagemenator:
 3: import have, itil.Has.lSe ;
 4: import availtil.Set;
 &: import veks.c.re.Marrix:
 9: * A class to merge groups (connected components) into a groups
10: * by using the feature strenaths between their nodes.
Tit it deprenaned
12: * Wauthor Jamie
13: *
14: **
15: public class Morger
A 1
1.75
       public static Set<Set<Integer>> mergeGroups(Set<Set<Integer>> groups,
1 23 :
            Marrix featureStrengthMx, int n)
100
20:
            // We only want n groups, therefore we need
31.
            is to merge groups in the current list until
            // there is a sufficiently small number.
27:
23:
24:
            // To pick which pair of groups to merge, we
25:
            // find the maximum inter-group bond strength,
26:
            // by taking the average of all the element pairs
27:
            // in the cross product.
28:
Z9:
            fcatureStrengthMx = MatrixUtils.normaliseToMax(featureStrengthMx);
30:
3~:
            // First, we remove any groups that only have one element. i.e.
32:
            // it has no convection with anything.
33:
            Set<Set<Integer>> groupsCopy =
               (Set<Set<Integer>>) ((HashSet<Set<Integer>>) groups).clone();
34:
35:
            for(Set<Integer> group : groups)
35:
               if(g:sup.slze() = ''g:supsCrpy.reneve(group';
3/1
            groups = groupsCopy;
38:
:0.
            if(teatureStrengthMx.numRows()!=featureStrengthMx.numCclumns())
               throw new Error ("Must have same number of rows as columns.");
40:
4
            System.out.println("Merging into "+n+" groups.");
42:
13:
46
             while 'grouns.size' >n'
45:
               System.oup.printin("Current group size: "-aroups.size());
46.
47:
               Setk_nteger> bestGroup1 = null;
48:
               Setkinteger> bestGroup2 = null;
               double pestPatrStore = 0.0;
1.5
57:
51:
               for (Set<Integer> gr. .. : groups)
52:
                    for(Set<.httpget> group2 : groups)
53:
                    [ if(!group'.equals(group/))
54:
55:
                             int linkCount - 0;
                             double correctPairStore = (...;
57:
                             for 'Integer : : aroup1'
58:
                             for(Integer ) : group2,
99:
60:
                                     chinectiair Scote+
61:
                                             Taurus MatsetytaMx.detrlement i, t ;
62:

 2. 28 % 20 % 5~+ 1.
```

```
double avgSours = varier=FairS iro
                   new DowbleddamkCountry
                if (avgSobre>restrainS (rw)
                   bestPairScore = avgScore;
                   pestGroup1 = gr up1;
                   bestGroup3 = group2;
       }
   // System.out.printin("Best pair score: "+bestPairScore);
   // System.aut.println("Best group !: "+bestGroup!);
   // System.out.println("Best group 2: "+bestGroup2);
   // Merge 2 groups:
   1/ 1. Add element of second to first.
  for(Integer = : bestGroup2)bestGroup1.add(1);
  // ?. Delete second from list.
  // System.out.println("Removing "abestGroup2+" from "agroups);
  // boolean removed groups.remove(bestGroup1);
   // System.out.println(removed);
  Set<Set<Integer>> newGroups = new HashSet<Set<Integer>>();
  for(Set<Integet> group : groups)
        if(!group.equals(bestGroup2))newGroups.add(group);
  groups = newGroups;
return groups;
```

schemagenerator/SchemaUtils.java

```
): package stmemauenerator:
                                                                                            o3:
                                                                                                            int == 0:
                                                                                            24:
                                                                                                            while briteudy (1)
 a: import tuva. ..P.ifferedPrader;
                                                                                            65:
 4: import java..o.FileReader:
                                                                                                                String line = pr.ipadbine();
                                                                                            cent.
 b: import favo. it 1.HashSet;
                                                                                                                if'i==n)f
 o: import fava.utal.!inkedfist;
                                                                                                                   list<String> appropriate = new LinkedDist<Strint> );
 import =ava.util.!ist;
                                                                                                                   // Reading attributes
                                                                                            69:
 H: import pava.util.Map;
                                                                                                                   line = line.supstrinu(', ine.langth, )-');
                                                                                            70:
 H: import [ava.util.Set;
                                                                                                                   String() lineralis : i.ue.split(", ";
 import java.util.TreeMap;
                                                                                                                   for(int !=0: _<!ineFarts.length; 5++)</pre>
                                                                                                                        attributes.add..ineFarts([]);
11: import nava.u*:1.Treefet;
                                                                                            11.3
                                                                                                                   if(!attributes.equals(attributesToCheck))
                                                                                            74:
13: 7**
                                                                                             /5:
                                                                                                                        throw new Error, "This mark scheme file is not " .
14: * Provides utilities with regards to reading mark scheme/list
                                                                                            76:
                                                                                                                            "compatible with the training model. " +
15: * files.
                                                                                            77.
                                                                                                                            "Ensure that their attributes are the same.");
16: * @author Jamie
                                                                                            78:
1/: *
                                                                                            79:
                                                                                                                + else (
18: */
                                                                                                                    // Remove [ ]
                                                                                            :08
19: public class SchemaUtils {
                                                                                                                    line = jine.substring(',.ine.length()-1);
                                                                                            81:
       public static void print (Set < Set < Integer >> groups,
                                                                                            82:
                                                                                                                   String[] lineParts = line.split(", ");
21:
             ListkString> featureNames)
                                                                                                                   Set < Integer > cluster = new HashSet < Integer > ();
                                                                                            83:
22:
                                                                                                                   for(int |=0; i<lineParts.length; j++)</pre>
                                                                                            84:
23:
             for (Set < Integer > group : groups)
                                                                                            85.
                                                                                                                        cluster.add(new Integer(lineParts[7]));
24:
                                                                                            86:
                                                                                                                   clusters.add(cluster);
25:
                System.out.print(("{ "));
                                                                                            87:
76:
                // Print each group
                                                                                            88:
27:
                int c . 0;
                                                                                            89:
                                                                                                                2++2
28:
                for (Integer 1 : group)
                                                                                            90:
                                                                                                            ł
29:
                                                                                            91:
30:
                     if(c>0)System.our.print(", ");
                                                                                            92:
                                                                                                            return clusters;
31:
                     System.out.print(featureNames.get(i));
                                                                                            93:
                                                                                                        } catch(Fxception e){
32:
                     C++;
                                                                                            94
                                                                                                           throw new Error("Problem reading mark scheme file");
33:
                                                                                            95.
34;
                System.out.print)n(" \"):
                                                                                            96
35:
                                                                                            97:
                                                                                                   ì
       ì
56:
                                                                                            98:
.17:
                                                                                            99:
                                                                                                   public static list<Set<Integer>> readMatkListFile(String inputFile,
38:
       public static void print(List<Set<Integer>> groups.
                                                                                           100:
                                                                                                        List<String> attributes, int numClusters)
ξQ -
             List<String> FeatureNames)
                                                                                           197:
                                                                                                        List<Set<Integet>> clusters = new linkedhist<Set<Integet>>();
40:
                                                                                           102:
                                                                                                        for(int n=0; i<numClusters; n-r) clusters.add(new HashSet<Integer>());
11:
             for (Set< integer > groups)
                                                                                           103:
                                                                                                        Map<String,Set<Integer>> groupMappings =
42:
                                                                                           102:
                                                                                                                SchemaUtils.getGroupMappings(inputFile);
73:
                System.out.print(("{ "));
                                                                                           105
44:
                // Print cach group
                                                                                           106:
                                                                                                        // We have to combine 'attributes' and 'groupMappings' to obtain
45:
                int c = 0;
                                                                                                        // 'clusters'. In the markinst file, groups numbers go from 1 to n,
                                                                                           1071
46:
                for(Integer i : group)
                                                                                           109:
                                                                                                        // whereas in lists we reference 0 to n-1, so we deduct 1
47:
                                                                                           119
                                                                                                        // accordingly.
48:
                     if(c>0)System.out.print(", ");
                                                                                           110:
                                                                                                        for(int i=0; i<attriputes.size(); i--)</pre>
40%
                     System.out.orint(reatureNames.get(1));
                                                                                           111:
50:
                                                                                           112:
                                                                                                           Set<Integer> group*appings.gef(attributes.get(i));
51:
                                                                                           113:
                                                                                                            for(Integer group : groups)clusters.get(group-1',add(1);
                System.out.printin(" }");
52:
                                                                                           114:
53:
                                                                                           115:
54:
                                                                                           116:
                                                                                                        return clusters;
public static past<Set<integer>> readMarkSchemeFile(String inputFile,
                                                                                           1:7:
56:
             wistkString> attributesToCheck)
                                                                                           1:8:
5/:
                                                                                           119:
5세:
             List<Set<'nreger>> plusters = new linked(ist<Bet<leteger>>();
                                                                                           120:
                                                                                                    * Produces a map of feature names to a last of proups
59:
                                                                                           171:
                                                                                                    * each corresponds to. It takes a .marklast fale as
cā:
             try
                                                                                           1/2:
                                                                                                    * ins impur.
ci:
                                                                                           1/3:
                Butforware when by a new ButforedRoader new FileReade: inputline, ;
ež:
                                                                                           1.4
                                                                                                   public static MagkString, Forkintegenes conformutMaggrads String importation
```

```
** 02/20/08<sup>*</sup>
```

15:36:06

schemagenerator/SchemaUtils.java

```
125: .
126:
            MackString, Set<Thteger>> groupMappings =
                    new TedeMapy Subura, Serkutions as a pro-
127:
125:
120:
139:
               Bursersakeades h = new BufseredReades(new Fileheades(inputFile));
131;
               while(pr.ready())
1.12:
                    Strong[] line | br.readline().split/"\t");
1237
134:
                     String attribute line[0];
135.
                    Boting() groups =
134:
                     line '[.substring(),)ine()].length()-1).split(", ");
                    groupMappings.put(attribute, new Treeder<Thteger>(,);
137:
                    if(!.ine[1].equals("[]"))
138:
136:
                       for(int | 0; :kgroups.length; :++)
                            groupMappings.get(attribute).edd(
140:
1411
                               new Integer(oroups|1;));
142:
143:
144:
               catch(Fxcaption e){
145:
               e.printStackPrace();
146:
             return groupMappings;
147:
148: }
149: )
```

schemagenerator/SetUtils.java

```
1: package schemagemenator;
3: import javaloril.HistSet;
4: import java.cril.Set;
5:
6: 2 **
7: * @deprecated
8: * @autnor Jamie
1.
10: 1/
11: public class SetUtils +
· 2:
      public static ShtkShtkTitteder>> romovhSubsets(SotkSetkInteger>> groups)
13:
14:
           // Clean up: merge aroups when one is a subset of the other.
15:
16:
           1 /:
           for (Set<Integer> s : groups)groupsTemp.add(s); // Make copy of groups
18:
19:
           for (Set<Integer> s : groups) {
20:
             for (Setkinteger> t : groups) ;
21:
                  if (isSubset(s, t)&&s!=t)
22:
                     groupsTemp.lomove(s);
23:
24:
25:
           return groupslemp;
26:
27:
28:
29:
      public static boolean isSubset(Set<Integer> s, Set<Integer> t) +
30:
           for (Integer 1 : s)
31:
              if (!t.contains(i))
32:
                  return false;
33:
           return true;
34:
35:
26:
```

14:33:27

package setantur;

semantic/SemanticModel.java

62:

```
r3:
                                                                                                              if swamels statement and
                                                                                          A4:
 4: import feva...; il.HashMap;
                                                                                          65:
                                                                                                                      svars.get(ref .addsted.symbol,p slyps ;
 =: import rava, n,il,litked.ist;
                                                                                                                else if harantt mull'
 5: import mava.ctul.List;
                                                                                         1.1 2
                                                                                                                      if(parent.getVar(ref)! null)parent.getVar(ref).addFred
                                                                                          67:
6: import gava. while Map;
                                                                                     | (Symbol, bosType);
 : import fava.attl.Sef;
                                                                                          4 ÷ :
                                                                                                                       else declareVar(ref):
8: import hava.util,TreeSet;
                                                                                          60:
                                                                                          70:
10: 11:
                                                                                                              else declareVarirefi:
                                                                                          11:
11: * This class is a representation of the semantic DRS
17: • expression, in object oriented form.
                                                                                         72:
                                                                                          73:
                                                                                                      Adds a pipary relation, of the form "symbol (ref), ref))".
13: * Its main method linearises the DRS into a set of
                                                                                          /4:
                                                                                                      public void addRel(String ref), String ref2, String sympol)
14: * semantic atoms.
                                                                                          15.0
15: * @acthor Jamie
16: *
                                                                                          <sup>1</sup>6:
                                                                                                              srels.add(new SemarticRel(ref1, ref2, symbol));
                                                                                          77:
17: */
                                                                                          /8:
18: public class Serant GModel {
            // A list of variables in the DRS.
                                                                                          19:
                                                                                                      // Adds an equality, of the form ref1=ref2 or ref11-ref2.
22:
            Map<string,SemanticVer> swars = new HashMap<String,SemanticVar>();
                                                                                          ម0:
                                                                                                      public void addEq(String ref1, String ref2, Boolean sense);
                                                                                          81:
                                                                                                              edRefL.add(r=[1);
21:
            // A list of relations in the DKS.
23:
            List<SemanticRel> sress = new LinkedList<SemanticRel>();
                                                                                          82:
                                                                                                              egRefR.add(ref?);
                                                                                                              eqSense.add(sense);
            // The LHS of equalities within the DRS.
                                                                                          83:
23:
            // The ith element corresponds to the ith element of eqRefR.
                                                                                          04:
24:
            List<String> eqRefl - new !inkedlist<String>();
                                                                                          85:
25:
                                                                                          86:
                                                                                                       / Sets the cardinality of a variable.
20:
            List<String> ogRefR new hinkedList<String>();
                                                                                          87:
27:
            // Whether each of the above are equalities of inequalities.
                                                                                                      public void setCard(String ref, int card, String cardType)
28:
            ListkBooleans egSense - new !inkedListkBooleans();
                                                                                          88:
                                                                                                              svars.get(ref).setCard(card); ;
29:
                                                                                          89:
                                                                                          90:
                                                                                                      public void changeSense(){
30:
            // The current sense while navigating the DRS tree.
                                                                                                                                       sense = !sense; }
                                                                                          91:
                                                                                                      public boolean getSense();
3 : :
            Boolean sense = true;
                                                                                                                                       return sense; }
                                                                                          92:
            // The current parent while navigating the DRS tree.
32:
                                                                                          93:
                                                                                                      public void setName(String ref, String symbol, String neType)
53:
            // null is at the root.
                                                                                          94:
            SemanticModel parent:
                                                                                          95:
35:
                                                                                                              svars.get(ref).setName(symbol);
            .. The set of generated semantic atoms from the model.
                                                                                          96:
ie:
            Set<String> aloms = new TreeSet<String>();
                                                                                          97:
37:
                                                                                          98:
                                                                                                      // Used for debugging purposes.
38:
39:
                                                                                         99:
                                                                                                      // Prints the variables and relations from
            // Methods to add atoms to the growing set.
                                                                                         100:
                                                                                                      // the tree.
            public void addAtom(String atom)[
                                                   atoms.add(atom);
40:
                                                                                         101:
                                                                                                      public void print()
            public void addAtcms(SetkString) aforsTeAdd'; 4'ers.addAll(atomsTeAd
41:
                                                                                         102:
                                                                                         1:2:
42:
                                                                                                              for(SemanticVar syar : syars.values())
             / Cets a variable given its name.
                                                                                         104:
43:
                                                                                         165:
            public SemanticVar getVar(String ref)
                                                                                                                      svar.pr.mi();
44:
                                                                                         106:
45:
                                                                                         107:
46:
                    SemanticVax sv = svars.get(ref);
                                                                                                              System.out.println("*** Relations ****");
                    // The var might be located in the varent model is not local.
                                                                                         :38:
47:
                                                                                         109:
                    if(sv:=null return sv;
                                                                                                              for(SemanticRel srel : srels)
18:
                    else return detParent(,.getVar(ref);
                                                                                        110:
(.
                                                                                                                       System.put.putmils.srel.tcString());
                                                                                         . . . .
50:
57:
                                                                                         172:
                                                                                         1132
                                                                                                      // Using the information accumulated from the tree,
            // Declares a variable.
52:
                                                                                        114:
            public void declareVar(String ref)
                                                                                                      // a list of semantic atoms is generated.
53:
                                                                                        115:
                                                                                                      public Set<String> getAiors()
74:
55:
                    svars.out(ref, new Semant(cVar(cer));
                                                                                         16:
                                                                                         117:
-6:
                                                                                                              for(SemanticVar svar : svars.va)(les())
                                                                                        115:
                                                                                         -19:
58:
            // Adds a predicate, where ret is the name of the variable,
                                                                                                                      n vai.defA! htts';
-9:
               sympol is the grazy relation being applied to it, and
                                                                                         120:
            // posType is the type of relation.
                                                                                          2.1
                                                                                                              for 3 partitives about are s'
e-2:
                                                                                        12::
            public void adultationications of, a sure of 1,20 in a realyne
```

semantic/SemanticModel.java

```
2::
                               srel.getAtoms(this);
 24:
 ٠<u>٤</u>٠:
 26:
                         Deal with equalities.
 2::
                      for int 1 0; ikeaReflisize('; i++'
 25:
 129:
                               SemantibVar 1 = getVar(eqRefil.get(t));
 135:
                               JemanticVar r = getVar(egkefR.get i)';
٠ ; :
                               String senseStringStart = "";
 13/:
                               String senseStringFnd = "";
 133:
                               if(!eqSense.get (1))
 134:
 135:
                                       senseStringStart = "![";
 130:
                                      senseStringEnd = "]";
 13/:
 138:
                               for(String atomL : 1.getAtoms(this))
 139:
                                      for(String atomR : r.getAtoms(this))
 140:
                                               // - is commutative.
141:
                                               atoms.add(senseStringStart + atomi+" =
"+atomR + senseStringEnd);
142:
                                               atoms.add(senseStringStart + atomR." =
"+a omT + senseStringEnd);
143:
144:
145:
                              1++;
146:
 147:
 148:
                      return atoms;
 149:
150:
151:
              public ScmanticModel getParent() {
 152:
                      return parent;
 153:
 154:
 155:
              public void setParent(SemanticModel parent) {
 156:
                      this.parent = parent;
15/:
158: J
```

```
02/14/08 14:10:09 1
```

semantic/SemanticRel.java

62:

63: 64:

ან:

56:

6/:

98: 60:

70:

7,2;

73:

745

75:

76: 77:

```
i: package somentin;
   u: import taza... 1.Eusnäet;
    1: import java. mul.Set;
   e: /**
   7: * Represents a relation in the semantic model.
   8: * dawthor Jamue
   9: *
   10: */
   it: public class SamantieFel (
   ۷.
  13:
               // Represents:
   4 •
              // harerrype(ref1,ref2)
   ^5:
              Strang refl:
   16:
               String refl;
   - - -
               String pare:
               String type = "";
   8:
   19:
   20:
               // Relations which we ignore.
   21:
               String[] rejectableKeywords = { "topic" ;
   22:
               // We may need to produce the list of atoms
   23:
               // numerous times, so they are cached for
   24:
               // subsequent use.
   25:
   25:
               Set < String > cachedAtoms;
   27:
   28:
               public SemanticRel(String refl, String refl, String symbol)
   29:
   35:
                       this.ref1 - ref1;
                       this.ret2 ref2;
   31:
                       this.name symbo);
   32:
   33:
   34:
   ან:
               public String toString()
   36:
   37:
                       return name="["rrefl","=refz="]";
   38:
   39:
               // The rule for getting atoms is as such:
   4U:
   4":
                / * We get the sets of atoms A and B for the
   42:
               // left and right variables in the relations.
   43:
               in i We take the sec:
   7.4
               / / name(a,b) / a<-A, b<-B /
   45:
               public Set<String> getAtams(SemantleModel mcdel,
   46:
   47:
                       if(cachedAtoms! null)return rachedAtoms;
   48:
                       $atk8thing> set = new Rash8et<String>(';
   40:
                       String senseString : "";
   50:
   ə1;
                       if (!model.getSense( )setseString = "!";
   52:
                       if(!shourd*sject())
   53:
   54:
   55:
                               SetkAtring> set5 = model.getVar(ref1).getAtons(model);
                               SockString> serR - mugel.uetVar.nei/'.detAtoms model);
   20
  57:
   58:
                               for (String atomb : setD)
   99:
                                       for Similar arons : seis'
                                               set.add(senseStr.nd)name="["-at mu-","
   823
+atams+"]"()
  7 2
```

semantic/SemanticVar.java



```
63:
 1: package someoute:
                                                                                          o4:
                                                                                                      •/ Gets a list of atoms represented by
                                                                                                         the variable.
 3: import havalar (MassSet)
                                                                                          ero:
                                                                                                      public Set<String> getAscas,SemanticModel madel*
 4: import fava.att...t.nkcd .st;
                                                                                          00:
 5: import favalunting and
                                                                                                              if magnedAtins!=null return cachedAtins:
 6: import fava. Fil.Sary
                                                                                          1112
                                                                                          :63
                                                                                                              String senseString "";
 a: public class Semant: Var -
                                                                                           ٠.:
                                                                                                              // CI the Idina
            / (att: 1 ^ ... ^ artrN) [name]
10:
                                                                                           12:
                                                                                                              Set<String> set - new HashSet<String>(.;
11:
            String ref:
                                                                                           / %:
                                                                                                              if(showldReject(name))return set;
12:
                                                                                           74:
            String name;
13:
                                                                                          75:
             List<String> attrs - new LinkedList<String>();
                                                                                                              if(name==null&&prop==null)
14:
                                                                                           76:
15:
             // Generic variable names which are ignored.
                                                                                          77:
16:
             String[] rejectableKeywords = [ "event", "proposition" };
                                                                                          78:
                                                                                                                      return set;
                                                                                          79:
17:
18:
                                                                                          80:
             String type = "":
                                                                                                              if(prop==null)
19:
             int cord = 1; // need to change to lower and upper bound
                                                                                          81:
20:
                                                                                          82:
                                                                                                                      // Get variable name itself
21:
                                                                                          83:
             SemanticModel prop;
                                                                                                                      set.add(senseString+name);
22:
            public void setFrep(SemanticModel model) {
                                                             prop = model; }
                                                                                          84:
                                                                                                                      // Get cardinality if present
23:
                                                                                          85:
                                                                                                                      if(card!=1)set.add(senseString="\"-;.ame,")="+card);
24:
             // We may need to produce the list of atoms
                                                                                          86:
                                                                                          87:
                                                                                                                      // Get var with attributes
25:
             // numerous times, so they are cached for
                                                                                                                      for(String attr : attrs)
                                                                                          88:
26:
             💚 subsequent use.
27:
             Set<String> cachedAtoms;
                                                                                          89:
                                                                                                                              set.add(senseString+att:+"["rname+"]"); // add
28:
29:
                                                                                        actr with subject
            public SemanticVar(String ref) { this.ref = ref; }
30:
                                                                                          91:
                                                                                                                              set.add(senseString+attr);
                                                                                                // add attr
31:
             // We the DRS centains propositions such as
32:
             // John(x1) and large(x1), we need to distinguish
                                                                                          92:
33:
                                                                                          93:
                                                                                                              ] else [
             // between the name of the variable, and attributes
                                                                                          94:
                                                                                                                       // We have a proposition
34:
             // relating to it. We can determine this by the
                                                                                                                      Set<String> childAtoms = prop.getAtoms();
                                                                                          95:
35:
                                                                                                                      // Add original child atoms to this level
                                                                                          96:
36:
            public void addFred(String symbol, String posTyce)
37:
                                                                                          97:
                                                                                                                      set.addAll(childAtoms);
                                                                                                                      // Add attributed child atoms if any attributes
38:
                     if(!shouldRenect(symbol))
                                                                                          98:
                                                                                          90:
                                                                                                                      for(String attr : attrs)
39:
                                                                                          100:
                                                                                                                              for (String atom : childAtoms)
4٦:
                            if(posTypeLequals("n")) ''posTypeLequals("v"))
                                                                                                                                      if(!snouldRefeat.but)
                                                                                         1.2:
42:
                                     rame : sympol;
                                                                                         1030
                                                                                                                                              sert alia (sersebit) no-arto-"["-a
43:
                             else
                                                                                       50m-"]");
44:
                                     attrs.add(symbol);
                                                                                                                                              set.aud(senseString-artr);
45:
                                                                                         104:
45:
                     · else {
                                                                                         105:
47:
                            type = symbol:
                                                                                         106:
48:
                                                                                         107:
49:
                                                                                         108:
                                                                                                              rodel.addAtoms(set);
50:
                                                                                         109:
                                                                                                              cachedAtoms = set;
51:
                                                                                         110:
                                                                                                              return set;
            // Sets the cardinality of the variable.
                                                                                         111:
52:
            public void setCard(int card){ this.card card;
53:
                                                                                         112:
                                                                                                      public String getName() [
54:
             // For debug nurposes.
                                                                                         113:
55:
            public void print('
                                                                                          114:
                                                                                                              return name;
                                                                                          ...
56:
                                                                                          116:
                     System.out.organizatin' "**** "-ret." *****" ;
                                                                                                      public void settame Strong name ()
                    System.out.println name ;
                                                                                          · 4 .
                                                                                                              this, the - tame;
53:
                    for Presidents: attes
                            Bystem.put.compating attra "[".rame."]".;
. :
                                                                                           7:
. .
                                                                                           21:
. 2:
                                                                                         2 :
                                                                                                      public strops of yre'
```

semantic/SemanticVar.java

```
122:
                    return type;
123:
· 2 ::
            public void selType(String type) (
125:
                    this.l/pe = '/pe;
127:
128:
            public int getCard(' !
129:
130:
                    return card;
731:
2:
133:
             private boolean shouldRepent(String symbol)
134:
135:
                     for(int u=9; _krejestableKeywords.length; i++)
136:
- 17:
                            if(rejectableKeywords[]].equals(symbol))return true;
138:
139:
                     return false;
140:
141:
1/12: 1
```

```
is package scelling;
    .: import =3vu.in.b..of-reakeade*;
    4: import java.io.P.laReager;
    6: import | java.illEntitWriter;
    6: import fava.util.HashMap;
    7: import java.utii.Map;
    3 / TA
   10: * Takes a CSV containing the English sentences and
   ": * "spell correction mapping file" (where each line
   12: * contains a spelling correction), and produces a
   13: * new CSV with the corrections made.
   14: * @author Jamie
   15: +
   6: 1/
   i/: public class ManCorrections (
               final static Strang noatFile = "./files/13bit-answers.csv";
               final static String outputFile = "./files/13bii-answers2.csv";
   19:
               final static String mappingFile "./files/13bii-spelling-corrections.
   4Û:
   21:
               Map<String,String> correct.ons = new HashMap<String,String>(;;
   22:
   23:
               public static void main(String[] args) {
   24:
   25:
                       MapCorrections by = new MapCorrections();
   26:
                       if(aros.length==0)mc.mao(mappingFile,inputFile,ourgurgile);
   27:
                       else no.map(args'0,,args[1],args[2]);
   28:
   29:
   30:
               public void map(String mappingFile, String isputFile, String outputFil
e)
   31:
   32:
                       try
   J3:
                                  Read in spelling corrections
   4
                               BufferedReador br - new BufferedReader (new FileReader,
   35:
mappingFile));
   36:
                               while(b).ready())
   37:
   38:
                                       String[] parts = br.readLine().split("\t");
   39:
                                       if (parts, length! /) System.out.print(n."Problem
 with line "-parts 11,7
                                       corrections.put(parts[0], parts[1]);
   40
   41:
   42:
                               br. blose();
   43:
   44:
                                . Read in imput file, and for each line, make the cor
rections
   45:
                               // and write to output file.
                               br = new BifteredReader(new FileReader(inputFile));
   40:
   47:
                               PrintWriter pw new PrintWriter(outputFile.;
   46:
   40.5
                               while(pr.ready())pw.println(transformLine(br.readLine()))
1));:
   50:
   51;
                               pr.close(.:
   52:
                               gw.pluse..;
   53:
   54:
                       ¹ catch ax moption e) i
   553
                               eut donáteckimane, p
```

```
57:
    - 2 :
    90:
                 public Strang transforming (Strang the
    ....
    1:0
                         for String key : corrections.keyper )
    ńź:
                                 if while to hower laws . .index if key > 0:
    03:
    64:
                                          int pos = line.toLowerlase().indexOf(key);
    55:
    56:
                                          // Need to check it's not part of some other w
 ord
                                          if(pos>0&&pos<line.rength()-key.length())</pre>
    58:
                                                  if (Character.isLetter(line.charAt pes
  1))&&!Chara:tor.isLatter(line.charAt(pos+key.longtn())))
                                                           " Not part of a word, so repl
    70:
 ace!
    71:
                                                          line : line.substring(0,pcs)->
 orrections.get(key)+line.substring(pos-key.length('';
    75:
                                          y else if (pos =0) t
    74
                                                  if (!Character..sLetter(line.charAt(pos
 +key. 'ength())))
    73:
                                                          line = corrections.get(key)+lu
| ne.substring(pos+key.length());
    78:
                                          | else {
    79:
                                                  if (!Craracter.isLetter(line.charAt(pos
 = 1 3 X }
    80:
    8...
                                                          // Not part of a word, so repl
 acc!
    82:
                                                          line = line.substring(0,pos)-
 errections.det(kev);
    83:
    84:
    85:
    36:
    8/:
                         return line:
    89:
    90. .
```

spelling/SpellCorrectionGenerator.java

```
1: package spelling:
                                                                                              5€:
                                                                                                                  for String val : wirdCounts.ketSet
    2: import __val___.5.f(eredRel.get)
                                                                                              57:
                                                                                                                             Correct plurality errors rock as "babys -> cap co".
    3: import hava.no.Fr@eReader:
                                                                                              - â :
                                                                                              50:
                                                                                                                          boolean correctedPlara. (y - spellUtils.correctFlinal)
    4: import havalto.Printwrites:
    5: import java.util.Map:
                                                                                           tyPrnone(
    o: import java.uril.TreeMap;
                                                                                              60:
                                                                                                                                         spelingCorrections, val):
    7:
                                                                                              61:
    8: import com.swabunga.spell.engine.Word;
                                                                                              62:
                                                                                                                          if (!corrected?lurality) {
                                                                                                                                  ... is the word in the dictionary?
    એ: import com.swabunga.spell.event.StringWordTokenizer;
                                                                                              63:
                                                                                                                                  boolean correct suel.Utils.isCorrect(val);
                                                                                              54:
   ~~: / * *
                                                                                                                                  // Is the word before the 's (if applicable) o
                                                                                              65:
   12: * Pases a dictionary as well as the original English
                                                                                           orrect?
                                                                                                                                  boolean correctA - spy. Utils.isAbostrophedw -
   13: * Inswers, and produces a "spell correction mapping
                                                                                              ėć:
       * file" which indicates what words should be replaced
                                                                                           dCorrect(val);
   15: * and what to replace them with.
                                                                                              67.
                                                                                                                                  if (correct II correctA I: val.length() <= .?) .
   16: * Outputting this correction file rather than modifying
                                                                                              68:
                                                                                                                                         // Leave short or correctly sociled wo
   17: * the CSV immediately allows manual alterations to be
                                                                                           rds alone.
                                                                                                                                         System.out.println("** " + val + " ->
   18: * made.
                                                                                              69:
   19: * @author Jamie
                                                                                           LEAVE"):
                                                                                                                                  } else {
   20: *
                                                                                              70:
   21: */
                                                                                              71:
                                                                                                                                          // We have a word not in the dictionar
   22: public class SpellCorrectionGenerator
                                                                                           y.
                                                                                              72:
                                                                                                                                          // If it's commonly occurring in the d
   24:
                // The dictionary.
                                                                                              73:
   25:
                final static String cicFile = "C:/Users/Jamie/Desktop/Programming/exam
                                                                                           ata-set,
marking/dictionary/eng com.dic":
                                                                                                                                          // leave it alone.
                                                                                              74:
                                                                                              75:
                                                                                                                                         if (wordCounts.get(val) > 5) {
   26:
                // The CSV containing optentially mispelled English answers.
   27:
                final static String fileToCheck = "./files/13bii-answers.csv";
                                                                                              76:
                                                                                                                                                  // Do nothing.
   28:
                // The file to cutput.
                                                                                              77:
                                                                                                                                          } else {
   29:
                final static String mappingFile = "./files/13bii-spelling-corrections.
                                                                                              78:
txt":
                                                                                                                                                  // SpellUtils.split(val) is no
                                                                                              79:
   30.4
                                                                                            longer used.
   33:
                                                                                                                                                  // and simply returns mill.
               public static void main (String[] args) throws Exception (
                                                                                              80:
                                                                                                                                                 String solitWord = spellUrils.
   32:
   33:
                        SpellCorrect.onGenerator scg = new SpellCorrectionGenerator();
                                                                                        split(val):
   3<:
                                                                                                                                                  if'sp.::Word!=null}
                        if(args.lengt): "fisog.generate'diffile.file10Thesk.mappingFile
1:
   35:
                                                                                                                                                          spellingCorrections.pu
                        else scg.generate(dicFile,aros[0),args[1);;
                                                                                              84:
   36:
                                                                                          t(val, splitword);
   37:
                                                                                              85:
                                                                                                                                                   else ·
                                                                                                                                                          double pestScore = Demi
   38:
               public void generate(String dieFile, String fileToCheck, String mappin
gFile) throws Exception
                                                                                          ble.NEGATIVE INFINITY:
                                                                                                                                                          String bestword = "";
   39:
                                                                                              8/:
   40:
                        // A utility class to check spellings.
                                                                                              88:
                                                                                                                                                          // Find the word in th
   41:
                        SpellUtils spe.!Utils = new SpellUtils(dicFile);
   42:
                        dufferedReager pr = new BufferedReadon(new FileReader(fileToCh
                                                                                          e data-set which
                                                                                                                                                          // has the strongest r
eakii;
                                                                                              90:
   43:
                                                                                           clationship w.'h
                                                                                                                                                          the misspelled word
   44:
                        // We keep a count of each of the words in the data-set.
                                                                                              9°:
   40:
                        Inis allows us to detect correctly spelled words which
                                                                                                                                                          for (String vall) : wir
   46:
                        If are not in the dictionary.
                                                                                              92:
   47:
                        Map<String, integer> wordCounts = to :areWords(IlleT)Inelk);
                                                                                           oCounts.kevSet(): -
                                                                                                                                                                  int editDistan
   48:
                                                                                              93:
   49:
                        // Remove common words
                                                                                           ce = StringUtils
   50:
                        wordCounts.comeve("the"):
   51:
                        wordCounts.remove("has");
                                                                                           .getLevenshtermDistance(val, val2);
                                                                                                                                                                  int letter...
   52:
                                                                                              95:
   53:
                        // Want to use wordCounts to produce a mar for alsopelled word.
                                                                                             StringUtila
   54:
                       MackString, string> spellingCtrie.t rs = new TreeMatkString,
                                                                                           .detletterant introduce val, val2 ;
                                                                                                                                                                  int : . . . . .
:lrint/ ;
                                                                                           rectus = Sterrittells
```

```
02/15/08
                                                            spelling/SpellCorrectionGenerator.java
                                                                                         .getDetherSetDufferenterval, val21;
.metPirstLotterkon(avval, va 1 :
                                                                                                                                                                 int firstDette
                                                                       int stroburt
                                                                                       - riomas - StringStria
 SimingUtals.derShael.Efercade(val,
                                                                                          134:
                                                                                        . perforation the Burgs/val. val2);
vai2 :
                                                                                                                                                                 int suzobliti
                                                                                         StingStris.getStzeDitierence, val.
                                                                       double share
                                                                                          13. :
  102:
                                                                                       \perp_{valz};
 (Marp. og (worddounts.get.valz,
                                                                                       1 13/:
   1.3
                                                                                          136:
.100/231
                                                                                                                                                                 double seeme =
                                                                                         1 (2 * editDistance)
  164:
- (3 * fars-letterBoxus)
                                                                                          139:
                                                                                       | - (3 * firstletterBonus) - JetterDiff
  105:
                                                                                       140:
 2
  106:
                                                                                       | - sizeDiff;
                                                                                           41.
* ed:tDis*ance
                                                                                                                                                                 if (score > De
                                                                                        stScore) +
  107:
- letterDiff sizeD:ff;
                                                                                        142:
                                                                                                                                                                          pestSc
  108:
                                                                       if (score > be
                                                                                       | ure = score;
stScore
                                                                                          43:
                                                                                                                                                                         bestWo
  109
                                                                                       I rd = val2:
&& !val2.equals(val)
                                                                                          144:
                                                                                          45.
                                                                                          146:
&& cditDistance <= 5 // We limit edit
                                                                                           141.
                                                                                                                                                         swellingCorrections.pu
                                                                                         t(val, bestWord);
                                                // distance
  112:
                                                                                          148:
&& wordCounts.get(val2) > wordCounts
                                                                                          149:
                                                                                           150:
                                                                                                                               ì
  113:
                                                                                          151 -
                .get(val) // Can only map to a
  1 4:
                                                                                          152:
                                                                                           153:
                                       // more popular word.
                                                                                          154:
  115:
                                                                                                                // Adds common spelling mistakes.
                                                                                                                sperlUtils.addCommonMistakes(sperlingCotroctions);
                                        // Avoids cycles.
                                                                                          155:
                                                                      1 1
  116.
                                                                                           156:
  157:
                                                                               bestSc
                                                                                           157:
                                                                                                                // Write correction mapping to a file.
                                                                                           158:
                                                                                                                PrintWriter pw = new PrintWriter(mappingFile);
ore = score;
                                                                              ревтМо
                                                                                           159:
  118:
                                                                                           160:
                                                                                                                for (String val : spollingCorrections.keySet()) ;
rd = val2:
                                                                                                                        System.out.print: "** " · val);
                                                                                          161:
  19:
  120:
                                                                                           162:
                                                                                                                        String lastVal = null;
                                                                                           163:
  121:
                                                              // Additionally find a
  122:
                                                                                          164:
                                                                                                                        // A word may map to a word which maps to another.
                                                                                          165:
                                                                                                                        String temp spellingCorrections.gel(val);
 word in the distinuary
                                                                                           66:
 12∄:
                                                              v. which has the closs
                                                                                                                        while (temp !: null) {
                                                                                          167:
t relationship with the
                                                                                                                                lastVal = temc:
                                                              // word. This may ever |
                                                                                          168:
                                                                                                                                System.out.print(" -> " - temp);
                                                                                           69:
ride a selection from
                                                                                                                                // Need to impore logarity mappings
                                                                / the data-set.
                                                                                           170:
                                                                                                                                if(!temp.equais(spe):ingCorrections.get(temp))
  125:
                                                              for (Object vai3 : spe | )temp = spellingCorrections.get(temp);
  26:
11Utals.petSaggestions(val)) {
                                                                                           171:
                                                                                                                                else tomo = null;
                                                                                          132:
                                                                      String val2 =
                                                                                          173:
  126:
                                                                                          114:
('Ward' vall', toStrino();
                                                                                                                        if lastVal! null'ow.orintin(val="\t"=lastVal);
                                                                                           75:
                                                                       int edutDistan
  729:
                                                                                                                        System.out.println();
                                                                                          176:
de - Stiningotyis
                                                                                          1771
                                                                                                                iw. Lose ';
                                                                                          ---
.detLevenshte.nDistance(val. valz);
                                                                                          .79:
  131:
                                                                       int letterDiff
                                                                                           8.:
   StringUni s
                                                                                                         * At him's for all the words in the data set.
  132:
                                                                                           181:
```

02/15/08 10:39:25

spelling/SpellCorrectionGenerator.java

```
132
  133:
               public static Matkanisha, Integery cyllaneW rds.Shrind fileNeto -
  184:
                       Magk Strong, Independ word Spants = new SteeMapk Strong, Intribute
; ;
  185:
  186:
                       try
  187:
                               bufleredReader br = new BufferedReader(new FileReader(
fileToCheck));
  lóó:
                               while (or.:eady()) (
  189:
  19.:
                                       String line - prichadtina();
  131:
                                       if(:=ne.:=ndex0f("\t")>")
  192:
  193:
                                               String semronno = line.split("\t")[1];
  194:
                                               StringWordTokenazer swt = new StringWo
rdTokenizer(sentence);
  195:
                                               while (swt.hasMoreWords()) {
  196:
                                                       String word = swt.nextWord().t
oLowerCase();
  19/:
  198:
                                                        if (wordCounts.get(word) = nu
11)
  199:
                                                               wordCounts.put(word, 0
);
  200:
                                                       wordCounts.put(word, wordCount
siges(word) + 1);
  201:
  202:
  203:
  204:
  205:
  205:
                       ] catch (Exception e) [
  207;
                               e.printStackTrace();
  208:
  209:
  2.0:
                       Systemio .t.pr ntin("*** Words ***");
  Ź .
                       for:String s : wordCounts.keySet();
  212:
                               System.out.println(s+" ("+wordCounts.get(s;+")");
                       System.out.printin("************);
  213:
  214:
  2:5:
                       return wordCounts;
 216:
 217:
 218: 1
```

02/15/08 10:43:50

spelling/SpellUtils.java

```
1: package spelling;
                                                                                          03:
                                                                                          F 4:
 /: import navaliatel e;
as import havaluutDExcept on;
                                                                                         +5:
 4: import mava.util. Tist;
                                                                                          င်ခန
                                                                                                        * Makes corrections with regards to incorrect suffixes
 ': import nava.util.Map;
                                                                                         . -:
                                                                                                        * for ringals. For example "babys" and "halls".
                                                                                          58:
 import god.swapinda.spell.erdine.sgellDintionary;
                                                                                                      public boolean poinesthiuma.styErrors MagkStitig, Strings spells.preat_
                                                                                          69:
8: import com.swabunga.spell.engine.SpellDictionaryHashMaD;
                                                                                      ons, String val.
9: import com.swapunga.spell.event.SpellCheckEyent;
                                                                                          70:
10: import com.swapunga.spell.event.SpellChecklisterer:
                                                                                         71:
                                                                                                               String transformedPhiral = "";
                                                                                                               if (var.engsW) th("ys")) transformedFloral = val.subsrcing(0, val.
if: import rom, swab.inga.scell.evant.SpeliChecker;
                                                                                          72:
12: import rem.swabunda.spell.event.StringWordTokenizez;
                                                                                       length:() 2. - "ies";
                                                                                                               if(val.endsW th("lfs"))transformedplural = val.substring(0,val)
14: /**
                                                                                       .length(3:3) + "lves":
15: * A wrapper to make the methods of a dictionary
                                                                                          74:
16: * labrary used more accessable. The necessaty is
                                                                                          75:
                                                                                                               if(!rranstormedFlura!.equals(""))
17: * mainly because the library has no direct method
                                                                                          /6:
18: * of determining a word's correctness, and alerts
                                                                                          77:
                                                                                                                       if(isCorrect(transformedPlural))
'9; * a SpeliCheckListener when a misspelling is
                                                                                          78:
21: * detected.
                                                                                          79:
                                                                                                                               spellCorrections.put(val, transformedPlural);
?': * @author Jamie
                                                                                          40:
                                                                                                                               return true;
22: *
                                                                                          81:
                                                                                                                       } else {
23: */
                                                                                          82:
                                                                                                                               return false;
74: public class SpellCtals (
                                                                                          83:
25:
                                                                                          84:
                                                                                                                else
26:
            boolean flag:
                                                                                          8⊃:
                                                                                                                       return false:
27:
            List suggestions;
                                                                                          86:
28:
            SpellD.ctionary dictionary;
                                                                                          87:
                                                                                                      j
29:
            SpeliChecker spellChecker;
                                                                                          Sc.
J0:
                                                                                          8'
                                                                                                      public void addCommonMistakes(Map<String,String> spellCorrections)
31:
            public SpellUttis (String fileName) throws ICException
                                                                                          90:
                                                                                         91:
                                                                                                               speliCorrections.put("aswell", "as well");
32:
33:
                    this (new SpeliDict) on arviash Map (new File (fileName));;
                                                                                          9/:
                                                                                                               speliforrections.put("dont", "don't");
                                                                                          93:
34:
                                                                                                               spellCorrections.put("everytime", "every time");
                                                                                          94:
                                                                                                               spellCorrections.put("colour", "colour");
35:
                                                                                                                                                                 // colour not
                                                                                       in dictionary!
36:
            public Speliffuls (Spelificationary dictionary)
37:
                                                                                         95;
                                                                                                               spellCorrections.pit("color", "colour");
                    this dictionary - dictionary;
                                                                                          96:
38
                                                                                                               spellCorrections.put("childs", "children"):
                    speli@hecker = new SpeliChacker(dist)shary);
                                                                                          97:
39
                                                                                                               spellCorrections.put("i", "I");
                spel.Checker.addSpellCheckListener(new SuguestionListener());
                                                                                          98:
40:
                                                                                          99:
41:
                                                                                         100:
                                                                                                      public boolean isApostropsedWordCorrect(String word)
17:
                                                                                         101:
43:
            public class SuggestionDistance implements age 1CheckListoner
44:
                                                                                         102:
                                                                                                               // Note, if isCorrect for the whole word returns true,
45:
                    public void spel_ingError(Spel!CheckEvont event) {
                                                                                         ຳບິວ:
                                                                                                               // then the output of this method does not matter.
                                                                                         104:
46:
                            flac = false;
                                                                                         105:
47
                            suggestions = eventiget5egg atlons,';
                                                                                                               int aIndex = word.indexOf("'");
461
                                                                                         106:
                                                                                                               // if it doesn't contain apostiophe, test not applicable.
٠٤.
                                                                                         107:
                                                                                                               if thinks return false;
50:
                                                                                         0€:
51:
                                                                                         109:
                                                                                                               if(word.endsWith("'s"))
92:
            public boolean isCorrect (String word)
                                                                                         110:
                                                                                         111;
53:
                                                                                                                       String sabword = word.substring([,aladex);
                                                                                         113:
54:
                    flac=true:
                                                                                                                       if(isCorrect(subword))return true;
                                                                                         . 13:
55:
                    spqluChecker.checkSpelling(new StringWithTikerLer(wird );
                                                                                                                       else return false;
                                                                                         114:
50:
                    return flag;
                                                                                                               } else {
5-:
                                                                                         115:
                                                                                                                       return false;
56:
                                                                                          5.
= .7
            public List getSuggestions(Strong word)
60:
                                                                                         118:
                    srel.Chocker.d.eckSpelichainew StaumawingT komuser(word);
                                                                                          ٠٠٠;
6 :
4 . :
                    return subsections;
                                                                                         'ZJ:
                                                                                                        ್ ಗಡೆಕ್ಟ್ರಾಂಡ್ ಕತ್ತಿ
```

02/15/08 10:43:50

spelling/SpellUtils.java

```
121:
                 * Used to detect if two words had been merged
  122:
                * together. sulic: "aswell") would return "as well"
                * and split "alot") return "a lot". It was lound
  123:
  24:
                 * the method led to poor results as it spir: words
  125:
                 * often when it shouldn't have.
  26:
  ._-:
               public String split String word.
  ∠d:
  774:
                       for ant x \ge 2; is word length: (x - 1)^{-1}; i.e.
  134: 77
 133: 7/
                               String wordL = word.substring(0,1);
 1.2: //
                               String wordR word.substring(1,word.length());
 133: //
                               if(isCorrect(wordL)&&isCorrect(wordR))return wordL+" "
+wordR;
 134: //
 135:
                       return null;
 136:
 137:
 138: J
```

spelling/StringUtils.java

ز4: 65:

6 · :

68:

72:

74:

75:

78:

82:

88:

90:

92:

03:

94:

95:

96:

Эв:

```
* * * *
5 2 B g
  ₹ <sup>64</sup> 9 14 3 1 1 3
```

```
: package speliing;
    a: import arg.unul.PagnSht;
    3: import = 3.3.0 1.301;
    L *
    6: * Provides comparison methods for strings.
    i: * @author Jamie
    8: *
    9: */
   40: public class StrangUtals &
    3:
               public static int getSizeDifference(String s, String t)
    3:
                       return Math.abs(s.length() - t.length());
   ٠.;
   (:
   7 :
               /**
   8:
                * Treats the words as a 'bay of letters', and
   9:
                * for 2 sets of letters A and B, computes:
   20:
                * #[(A^-B)v(B^-3)]
   21:
   22:
               public static int getLetterSctDifference(String s, String t)
   23:
   24:
                       Set<Character> sLetters = new HasnSet<Character>();
   25:
                       Set<Character> tLetters = new HashSet<Character>();
                       for(int i≈0; i<s.length(); i++)sletters.add(s.charAt(i));</pre>
   26:
   27:
                       for(int i≈0; i<r.length(); i*-'tLetters.add(t.charAt(i));</pre>
   28:
   29:
                       int diff ≃ 0;
                       for(Character c : sletters)if(!tfetters.contains(c))diff++;
   36:
                       for(Character c : tLetters)if(!sLetters.contains(c))diff++;
   31:
                                                                                        1 81;
   32:
   33:
                       return diff:
   34:
   35:
   3C:
   37:
                * As a heuristic, we class two words as particularly
   38:
                * similar of their first letter is the same.
   39:
                * Increfore "receive" and "recieve" would be more
                * similar than "believe" and "relieve".
   4J:
   41:
               public static int detFirstCetterBonus(String s, String t)
   42:
   43:
                       if(s.charAt(0) = -t.charAt(0))return (;
   44:
   45:
                       else return 0;
   46:
   47:
   48:
   49
                * String edit distance.
   50:
                * Unknown author.
   51:
   52:
               public static int detLevenshteinDistance (String s, String t, )
   53:
                          if (s == null | 1 t == null) [
   54:
                           throw new TilegalArgumentException."Strings must not be nu
11"1:
   55:
   56:
                         int r = s..ength'; Jength of s
   >7:
                         int = 1. ergth '; / length of t
   00:
   . 3
   . . :
                          if : =- 0) .
                           return ~;
```

```
62:
                          7 else if x => ' ]
    c3:
                            return 🕾
                          int c!! new int[:#']; //'previous' cost effay, norizontall
    67:
                          int ai] = new int(bri); // cost array, merizontally
                          int Lail: //placenoider to assist in swapping p and d
    69:
    7 C :
                          // indexes into strings s and t
                          int i; // iterales through s
                          int ; // ife: ates through t
     7.5.5
                          char t_j; // jth character of t
    76:
                          int cost; // cost
    17:
                          for (| ≈ 0; i<=n; i++) |
    19:
                             p[i] = i;
    80:
    81:
                          for (1 = 1; 1<=0; 1+4) 4
    83:
                             84:
                             d[0] = 0;
    85:
    86:
                             for (1 1; 1 < n; J = 1) {
                                cost s.cnarAt(i-1): at_j 2 1 : 1;
    87:
                                // minimum of cell to the left+1, to the top+1, diagon
I ally left and up +cost
    B9:
                                d[z] = Varh.min(Math.min(d[z-1]+1, p[z]+1), p[z-1]+co
                             }
    9::
                             // copy current distance counts to 'provious row' distance
 e counts
                             _d = p;
                             p = a;
                             d = d;
    97:
                          // our last action in the above loop was to switch d and p,
 SO D DOW
    Que:
                          // actually has the most recent cost counts
   100+
                          return p[n];
   101:
   102:
   133: 1
```

server/ClientCAndC.java

```
1: package server;
                                                                                              59:
    2: import rata, r. r;
                                                                                              F1:
    3: import pava.net.*;
                                                                                              52:
    4: import have.utal..Sch:
    5: import hava.utii.TheeSot;
                                                                                              63:
                                                                                              54:
    7: import warser.@RSreader:
                                                                                              65:
    9: public class ClientCAngC +
                                                                                              éé:
               final static int mort 4915); // port to which to connect
                                                                                              €7:
               final static String computer "curlew.comlab.ox.ac.uk";
   11:
               final static String endString = ".";
                                                                                              69.
   17:
   13:
                                                                                               1:
                                                                                              71:
   4
               PrintWriter toServer:
   15:
               BufferedReader fromServer;
                                                                                               2:
                                                                                              73:
   16:
   17:
               public static void main(String[] args) throws ICException [
                                                                                              74:
   8:
                                                                                              75:
                       BufferedReader in = new BufferedReader(new ImputStreamReader(3
                                                                                              76:
   19:
                                                                                              77:
vstex.in));
   20:
                                                                                              78:
   21:
                                                                                              79:
                       // initialize socket
   22:
                       System.out.printin("** Attempting to connect. **"!;
                                                                                              ao:
                                                                                              81:
   23:
                       Socket t = new Socket(compuler, purt);
                       System.out.println("** Client connected.");
   24:
                       System.out.println("** Please input a sentence to parse and bo
                                                                                              92:
   25:
x.");
                                                                                              83:
                                                                                              84:
   26:
   21:
                       // Associate writer with socker
                                                                                              85:
                       PrintWriter pw = new PrintWriter(t.getOutputStream(), true);
                                                                                              86:
   28:
                                                                                              87:
   29:
                                                                                              88:
   30:
                       // Also associate reader with socket for receiving back respon
                                                                                              89:
30
   31:
                       BufferedReader b: = new BufferedReader(new InputStreamReader(t
                                                                                              90: 1
                                       .detinputStream()));
   32:
   33:
   ...
                         A Reneatedly input string from terminal, and write to socket,
 until

    endString is imput

   35:
   36:
                       String st:
                       System.pur.print"> "";
   37:
   38:
                       st = in.readLine();
                       while lat.equals empSir up' [
   39:
   40:
                               pw.printlr.(st);
   41:
   42:
                                String response br.readLine();
                                System.out.println(response);
   43:
   44:
                                EntkStrings set = DBSreader.atemizerrospense);
   45:
                                for(String s : set)System.out.grintin(s);
   46:
   47:
   48:
                                System.ou..pijnt("> ");
                                st = im.readLine();
   49:
   50:
   51:
   52:
   - 3
                       System.cut.println("Connection closed");
   54:
               17
   5::
               public void comment. | throws IDException
   50.5
   ā:
                           This initial sorket
   56:
```

```
Spiker to new Spiket respicer, tiret ;
           Associate writer white look a
        toServer = new PrintWriter(tigetoutputStream(), true);
        // Also associate reader with socket for receiving back respon-
        fromServer = new ButforedReador(new InputStreamReador(+
                        uget The utrate an
        System.out.printls "**-- Connected to server":
public String getDRS(String sentence) throws IOException
        toServer.printin(sentence):
        String response = fromServer.readLine();
        System.out.println("*-- Got response from client.");
        return response;
public void close() throws 10Exception
        // Giving the sentence '.' tells the server thread to shut dow
        toServer.print c(".");
        // But we still need to close the connection.
        toServer.elase():
        fromServer.close();
        System.out.println("*-- Closed connection to server"):
```



server/ServerCAndC.java



```
1: package server;
    3: import gava.io.Buiterepkeadon;
    -: import java.is.FileReader;
    5: import java. Lo. InputStrear Reader;
    6: import java. G.PrintWriter;
    7: import java.net.ServerSocket;
    8: import java.net.Socket;
   10: /**
   11: * Listens to client requests to parse English sentences,
   '2: * and sends back the parsed expression in DRS form.
    * * Can operate over a remote connection.
   '4: * @author Jamie
   15: *
   16: */
   17: public class ServerCAndC [
   8:
   19:
                final static int port = 49157; // port to which to listen
   20:
                final static Spring INPUT_FILE = "input.txt";
   21:
                final static String INTERMEDIATE_FILE = "output.txt";
   22:
                final static String OUTFUT_FILE = "outputBOXED.txt";
   23:
   24:
                // Starts the soap server, which receives parse commands.
   25:
                final static String soapServerCommand = "../candc/candc-1.00/bin/soap_
server --server localhost:9000 --candc ../candc/candc-1.00/models/boxer";
                // Starts the soap client, which sends parse commands.
   27:
                final static String scapClientCommand = "../candc/candc-1.00/bin/scap_
client --input "(INPUT_FILE)" --output "+INTERMEDIATE_FILE+" --url http://localhost:90
00";
               // A command to 'box' the CCG to produce a DRS.
   28:
               final static String boxerCommand = "../candc/candc-1.00/bin/boxer --in
put "+INTERMEDIATE_FILF+" --output "+OUTPUT FILE+" --box true --flat":
   31:
               public static void main(String(] args)
   32:
                        Runtime r - Runtime.getRuntime();
   33:
                        try
   34:
                        Ą
   35:
   36:
                                ServerCAndC cle - new ServerCAndC();
   37:
                                execute(soapServerCommand, false, 20000); // Start the S
OAP server
   38:
                               atantServerListener();
   19
   40:
                        } catch(Exception e){
   41:
                                e.pr_ntStackTrace():
   42:
   43:
               3
   44:
   45
               public static void execute(String command, boolean wait, long delay)
   46:
   4 / :
                       try
   48:
   49:
                               // Execute the command via command line prompt.
   50.
                               Process p = Runtimg.getRuntime().exec(command);
   51:
                               // Some commands we can detect when they're finished.
   52:
                               // For starting the scap server, we don't know when it
   53;
                               // ready, so inscead we wait a suitable amount of time
   54:
                               if(wait)
   55:
```

```
p.waitFor();
   57:
                                 else i
                                         System.out.printin("Sleeping for ")ae'av+"ms t
   58:
o allow server to start,",;
   59:
                                        Thread.sleep(delay);
   60:
   6^:
                        ; catch(Fxgeption e){
   62:
                                e.printStack(race();
   63:
   64:
   65:
   66:
   67:
                 * Reads the file produced from boxing, and extracts
   68:
                 * the DRS string from it.
   69:
   70:
                public static String extractDRS()
   71:
   72:
                        try
   73:
                        4
                                BufferedReader br = new BufferedReader(new FileReader(
   74:
"./outputBOXED.txt"));
   75:
                                int 1=0;
                                String line "";
   76:
                                while(ik=13%&br.ready())
   77:
   78:
   79:
                                         line = br.rcadLine();
   80:
   81:
   82:
                                if (i 14)
   83:
   84:
   85:
                                         int JastBracketPos = line.lastIndexOf(')');
                                        line = line.substring(0,lastBracketPos).trim()
   86:
   8/:
                                        return line;
   88:
                                else
   89:
                                         // Case where the C&C parser failed
                                        return "";
   90:
   91:
   92:
   93:
                        } catch(Exception e)<sub>1</sub>
   94:
                                throw new Error(e);
   95:
   96:
   97:
   98:
                * Listens to incoming connection requests.
   99:
  100:
  101:
               public static void starfServer()stemer()
  102:
  103:
                        try
  104:
                                System.out.println("*** Listening for incoming connect
  105:
ions. ***");
  106:
  107:
                                // Initialise server socket
  108:
                                 ServerSocket Listener = new ServerSocket(port);
  100:
 110:
                                // We want the program to always accept connections, a
o we loop continuously.
  1113
                                while (true) [
 112:
```

```
02/14/08
12:04:36
```

server/ServerCAndC.java

```
- Wait for, and them accept, connections
114:
                            Porker a libreret annett.
                                       Create a new thread using this socket.
115:
1061
                            Borveringead of a new Jerver Targad & ;
                             f.start(): // Start the thread.
10:
159:
-20:
121:
                    | catch(%xception e){
-27:
                            g.printStackTrace();
123:
124:
.25:
126:
127:
             * The CAndC Parser reads the input string from a file.
128:
             * The method puts this string in a file to produce
129:
130:
             * the suitable input for the paiser.
191:
132:
             public static void writeinputToFile(String str)
153:
134:
                    try
135:
°3€:
                            PrintWriter pw = new PrintWriter(INPUT_FILE);
137:
                            pw.println(str);
128:
                            gw.slase();
139:
                    ; catch(Exception e) {
140:
                            e.printStackTrace();
141:
142:
143;
144: }
```

02/14/08 11:40:33

server/ServerThread.java

71: 3

```
1: package server;
    3: import _ava.io.BufferedReader;
    4: import java.io.InputStreamReader;
    5: import java.io.PrantWriter;
    6: import java.net.Socket;
    1:
    b: /**
    F: * A ServerThread in instantiated by ServerCAndC, and
   10: * represents a service for a particular client.
   11: * Several ServerThreads may be running if multiple
   12: * clients are being concurrently served
   13: * dauthor Jamie
   14: *
   15: */
   16: public class ServerThread extends Thread .
    181
               Socket s;
    9:
   20:
               public ServerThread(Scoket s){    this.s - s;
   2':
               public void run()
   22:
   23:
   24:
                   try
   25:
                       System.out.println("*** Accepted incoming connection. ***");
   26:
   27:
   ZB:
   23:
                       // Associate reader with socket
   30:
                           BufferedReader br =
   31:
                                     new BufferedReader(new InputStreamReader(s.getIn
putStream()));
   32:
   33:
                           PrintWriter pw - new PrintWriter(s.getOutputStream(), true)
   34:
   35:
   36:
                               // Read just 1 line
   37:
                               String st br.readLine();
   38:
    J9:
                               // Now read until we terminate.
   40:
                               while(!st.equals("."))
   41:
   42:
                                       System.out.println("Parsing: "-st);
   43:
   44:
                                       ServerCAndC.writeinputTcFile(s-);
   45:
                                       ServerSAndS.execute(ServerSAndC.shapS):encSamm
and, true, . ;
                parse the input using CandC
   46:
                                        ServerCAndC.execute;ServerCAndC.poxonCommand,t
rue,;;
                // bex the output CCG
   ۷٦:
                                       String drs = ServerCAndClextractCRS';
   48:
   49:
                                       System.out.print.n("DRS: "-ars);
   50:
   51:
                                       // Prints the DAS string to the sucket.
   54:
                                       pw.printin(ars);
   -3:
   54:
                                       System.out.orintln("*** Request successfully p
rocessed. ***":;
   55:
   , t
                                       st - br.readline();
   5/:
                                       if(st==null)break:
```

```
58:
59:
60:
61:
                             br.closet);
62:
                             pw.dlash();
63:
                              System.out.psin.t.... "*** Connection terminated ***");
64:
65:
                 } catch:Exception e)(
ьó:
67:
                     e.printStackTrade( ;
65:
69:
70:
```

```
02/20/08
14:25:29
```

pipeline/AnswerConverter.java

```
1: package pipeline;
    3: import | ava.ip.BuftereoReader;
    4: import java.i..FileBhadei;
    S: import gava.ip.PrintWriter:
   7: /**
    B: * Converts the pl file into a seater CSV.
   9: * Rows of the CV have the form:
   10: * "mark \: unswer"
   11: * @author Jamie
   ٠. *
   13: 47
   14: public class AnswerChaverier /
   15:
   ь.
         static final Strang impurrile = "./files/answers/2/4(d).pl";
         static final String outputhile - "./files/answers/2/4d-answers.csv";
   17:
   î.
   19:
         public static void main(String[] args) {
   20:
              try
   21:
   22:
                  BufferedReader br = new BufferedReader(new FileReader(inputFile));
   23:
                  PrintWriter pw : new PrintWriter(outputFile);
   24:
   25:
                  while(br.ready())
   76
   27:
                      String line = br.coadTine();
   23:
                      if(ine.indexOr("%") :=0:|line.indexOf("noanswer") > 0:|ine.ind
   9
ex0f("answer")!≈0)
   30:
   31:
                              // Reject
   32:
                      } else i
   33:
                              int openBrac = line.undexOf("(");
   34:
                              int nextComma = line.indexOf ",",openBrac);
                              char squre = getScore(ine,nextComma+1);
                              int openQuote = line.indexOf "\"", nexrComma);
   36.
                              int closeQuote = line.indexGi("\"",openQuote+1);
   37:
   38:
   30.
                              if(closeQuete>=0)answer = line.substring(openQuete+1,c
.oseQuote);
                              40:
   41:
                              answer = arewer.replaceAll("\t","");
   12:
                              dnawer - inswer.regladeAll.",",",";
   43:
   44.
                                      // make sure there's space before comma
                              answer = answer.ret_aceAll."\\."," \\.");
   .5.
   46:
                                      // make sure there's space before full stop
                              answer answer.forcareAll " "," ");
   47:
   48:
                                      // remove duplicate spaces
                              answer answer.ifu ;
   49:
   50:
                              System.out.printlh(score-"\t" answer);
                              pw.println(score+"\t"+answer);
   52:
   53:
   54:
   55:
   56:
                  ir.plose();
   57:
                 gw.u 585 7
   58:
   5₹:
                 catch.Exception of
                  e.pr/ntf/rackina/c();
```

pipeline/ArffGenerator.java



```
63:
                                                                                                               String line = br.readline();
 1: package pipeline;
                                                                                           64:
                                                                                                               if(line.indexCf("%"\!=0)
 2: import avi.io.BufferedReader;
                                                                                           65:
 3: import avaluate. LeRegue::
                                                                                                                       String() lineParts = 1 mo.split("\t" :
 4: import avalid.PrintStream;
                                                                                           66:
 5: import avaito.PrintWriter:
                                                                                           81:
                                                                                                                       int =neScore = new 's** per(_ineParts d]';
 f: import hava.util.HashMan:
                                                                                           o8:
                                                                                                                       if(theScare>maxScc.c)raxacore = theScare;
 7: import hava.util.linkedList:
                                                                                           69:
 8: import ravalutilings:
                                                                                                                       scores.add(theScore);
                                                                                           70:
                                                                                                                       Set(String> sor = DRAreider.otatize lineParts[1]);
 import java.uril.Map:
                                                                                           71 :
10: import hava.util.Seg;
                                                                                           72:
                                                                                                                       atomSets.acd(set;;
": import rava.::: TrecSet;
                                                                                           73:
                                                                                                                       ps.printin."Successfully parsed line "+numLines+".";
                                                                                           74:
'3: import parser.DRSreader;
                                                                                           75:
14:
                                                                                           76:
                                                                                                               puminoes++:
• 5 :
                                                                                           772
16: /**
                                                                                           78:
1:: * Generates an arff file.
                                                                                           79:
                                                                                                          br.close();
18: * The input file is a CSV with tab separated fields "score
                                                                                           60:
19: * It extracts all the semantic aroms from each DRS, collates them,
                                                                                           81:
                                                                                                          ps.println("Num lines: "+numlines);
20: * and then produces a binary table with the atoms as the columns.
                                                                                           82:
21: * @author Jamie
                                                                                           83:
                                                                                                          // Work out segments
22: *
                                                                                                          int startPos - 0:
                                                                                           84:
23: */
                                                                                           85:
                                                                                                          int segmentLength = 0;
24: public class ArfiGenerator f
                                                                                           86:
                                                                                                          int k=0;
                                                                                           87:
26:
       static String inputFile "./files/8b-answers-processed.csv";
                                                                                           88:
                                                                                                          while(k<=segnumeer)
27:
       static String outputFile = "./files/8b.arff";
                                                                                           89:
28:
                                                                                           90-
                                                                                                               startPos += segmentLength;
                                                                                                               segmentLength = numLines / numsequents; // takes divisor;
29:
       // If this is set to true, atoms which occur only 1 time
                                                                                           91:
30:
       // in the corpus are removed.
                                                                                           92:
                                                                                                               numsegments--;
31:
       final static boolean CONDENSE = true;
                                                                                           93:
                                                                                                               numLines -= seamentLength;
32:
       public static void main(String[] args)
                                                                                           94:
                                                                                                               k++:
33:
                                                                                           95:
34:
             System.out.println("Usage: ArffGenerator INPUTFILE.csv "+
                                                                                           96:
35:
                                                                                           97:
                                                                                                          if(!exclude){
                     "OUTPUTFILE.arff"):
36:
             generateFile(args[0],args[1],1,0,false,System.out);
                                                                                           98:
                                                                                                               startPos = -1;
3/:
                                                                                           99.
                                                                                                               segmentLength 0;
38:
                                                                                          100:
39:
       public static void generateFile(String inputFile, String outputFile,
                                                                                          101:
40:
                                                                                          102:
                                                                                                          // Want a set with all the atoms (but we want
             int numsegments, int segnumber, boolean exclude,
41:
             F: ntStream ps)
                                                                                          103:
                                                                                                          // to ignore the testing segment)
42:
       [
                                                                                          104:
                                                                                                          for(int i=0; i<atomSets.size(); i++)</pre>
43:
                                                                                          135:
                                                                                                               if(!(=>=startPos&&ikstartPus.scamenrLength):
44:
            os.orintin("EXECUTING ARFF GENERATOR"):
                                                                                          106:
45:
            ps.println("Input file: "-inputFile);
                                                                                          107:
46:
                                                                                         138:
                                                                                                                  Set<String> atoms = atomSets.get(1);
            ps.println("Output file: "+outputFile);
                                                                                          119:
                                                                                                                  for (String atom : atoms)
47:
                                                                                          **::
46:
             .st<.hteger> scores = new linkedList<Integer>(.;
                                                                                                                       if'a.'Atoms.get:atim =.null:allAtims.put(afor, 1 ;
49:
             ..st<Set<String>> atomSets = new linkedlist(Set<String>>();
                                                                                          112:
                                                                                                                       else allAtoms.puc(athr. allArams.get atom)="";
50:
             Mag<String,Integer> allAtoms = new HashMap<String,Integer>();
                                                                                          113:
51:
            Sot<String> finalAtomList = new TreeSet<String>();
                                                                                          1.4:
54:
                                                                                          115:
53:
            try
¬4:
                                                                                          116:
                                                                                                          // If an atom has a count of 1, we only include it if
                Bufferedweader pr = new BuffereaReader(
                                                                                          117:
o5:
                                                                                          118:
                                                                                                          // we're not condensing the list.
56:
                     new . : Lekeader (inputFile);
                                                                                          119:
                                                                                                          for(String atom : allAtoms.keySet().
5/;
                                                                                                               if(a,)Atoms.get(atum)>1)finalAtur...st.add(atom);
⊃8;
               int max3core = 0;
                                                                                          120:
                                                                                                                       else if(!CONDHNSF fibilArr (al.add:atom);
59:
               int rumingess;
                                                                                          121:
60:
                                                                                          27:
61:
               while 'hr.leady())
                                                                                          123:
                                                                                                          PrintWriter pw new PrintWrite: ( wij // lo);
62:
                                                                                          124:
```

pipeline/ArffGenerator.java



```
125:
16:
                if - k to 14
127:
                     pw.printlm "% This is part of a test suite.";;
25:
                     pw.printlin."% Took all lines except "ratarincar" to "-
129:
                             'srantEcs-segrentLergth-1 4".";
131:
                     pw.printin();
131:
132:
133:
4:
                gw.printin("@relation myexample");
35:
                for 'Siring afor : finalAtomList'
36:
                     pw.println("@attribute \""+atum="\" ( 0, 1 }",;
137:
138:
                String scoreString "@attribute score ( 0";
139:
                for(int i='; <=maxScore; i(+)sgornString+= ", "+);</pre>
140:
                scoteString+= " }";
141:
142:
                ow.println(scoreSpring);
43:
                pw.println();
144:
                ow.println("@data");
145:
146:
147:
                int 1 = \bar{0};
                for(Set<String> atomSet : atomSets)
148:
149:
                     if(!(1>=startPos&&i<startPos+segmentLongth))</pre>
150:
151:
                        for(String atom : [inalAtomiist]
752:
153:
                             // Check to see if the atomSet has this atom.
154:
                             if(atomSer.domtains(atom))pw.brin*("1");
155:
                             else pw.print("0");
156:
                             pw.pr.nt(", ");
157:
158:
159:
                        pw.println(scores.det(1));
160:
161:
162:
                     1++;
16::
164:
1651
                 pw.slose();
                ps.println("Successfully generated file "computFile;;
166:
16 :
             } catch(Exception e){
168:
                     e.printStackTrace();
1.49
170:
                     throw new Error ("Stopped");
----
1/2:
---:
```

pipeline/FileParser.java

```
1; package paperane;
 3: import nawa. na. A fine madication;
 4: import ava. c.Fi ekcaden;
 S: import (iva. to.Ph fr@friggh;
 e: import hava.id.StintWriter;
 b: import server.Client(AndC);
1:
1. . . * *
11: 1 Tukes a USV file of English answers and converts
12: * them into DRS semantic form using an external
î3: * pa∶ser.
14: * Mauthor Jamie
15: *
16: */
17: public class FileParser {
18:
19-
       static String imputFile = "./files/8b-answers.csv";
20:
       static String outputFile = "./files/8b-answers-processed.csv";
21:
22:
       public static void main(String[] args) {
23:
            processSentences(args[0], args[1], System.out);
24:
25:
26:
       **
27:
        * Processes a single sentence.
28:
29:
       public static String processSentence(String sentence)
30:
31:
            ClientCAndC client = new ClientCAndC();
32:
33:
34:
               client.connect();
35:
               String drsString = client.getDRS(sentence);
36:
               return drsString;
37:
38:
            } catch(Exception a){
39:
               throw new Error ("Failed to connect to server.");
40:
41:
42:
43:
       public static void processSentences(String imputFile,
44:
            String output: 11e, PrintStream bs)
45:
46:
            try {
47:
               os.or_atlan("*** Processing sentences. ***");
48:
               ClientCAsdC o'rent = new ClientCAmdC',;
49:
               orient.connecti::
5:
52:
               RofforedReader or = new BufferedReader new FileReader (inputFile.);
53:
               PrintWriter pw = new PrintWriter, outputFile);
54:
55:
               while (or. ready()) 4
56:
                    ps.println("Reading line");
                    String line br.readline();
58:
52:
                    int rextRos line.indexCi("\t") + 1:
£ 1:
                    String acore = line.substring(0, textPos - 1);
71:
6.):
                    String Bentance = Irro.substring(textPos);
```

```
if(!sertence.equais(""))
63:
04:
√8:
                            String dreString = 0 .ont.dec.ed.accompo :
                            if (!arsString.equals ""
co:
€7:
                               rs.println("!!! PARSED OK" ;
                               gw.println.sbcoc + "\t" +
                                    drastiing.replate "/"," or "::;
10:
                               priorintia "XXX FAILED");
12:
 3:
`4:
 ÷:
16:
77:
               os.printin("*** Finished processing sentences. ***");
78:
79:
80:
               br.close();
81:
               pw.close();
82;
83:
               client.close();
84:
            | catch (Exception e) {
85:
               e.printStackTrace();
86:
87:
88:
89:
90:
9::1
```

pipeline/KeywordIdentifier.java

```
6 :
 is package o restner
                                                                                                          e.printStackTrace();
 /: import -aval.o. bificredReade:;
                                                                                          . . :
 : import yava..c."ileReade:;
                                                                                          (5:
                                                                                                1
                                                                                          5.00
 4: import |ava.util.HashMap;
 i: import havaluti .Map;
 : import _ava.stl..SeT;
 /: import gava.uti..TreeSet;
 9: import parser.DRSreader;
1 :
15: 21*
's: • This identifies the most popular atoms in the dataset.
14: * Used for testing purposes only, and is not used in the
The " averal" pipeline.
16: * Gauthor Jamie
1 7 ± *
15: */
19: public class Keywordldentifier {
20:
       static String imputFule = "./files/5aii-answers-processed2.csv";
21:
37:
       public static void main(String(] args)
23:
24:
            if(args.length> 1)generateFile(args[0]);
            else generareFile(imputFile);
25:
26:
27:
25:
       public static void generateFile(String inputFile)
Zy:
30:
            Map<String,Integer> atomCounts = new HashMap<String,Integer>();
31:
32:
            try
33:
34:
               BufferedReader br = new BufferedReader(new FileReader(inputFile));
45:
               int numbines=0;
36:
               while 'br.ready!'
37:
30:
                    Strong line = bi.readLine();
39:
                    if(line.indexOf("%")! 5)
40:
                       String( lineFarts = line.splt("\t");
                       Set<String> set     DRSreader.atomize(lineParts[!]);
42:
                       for String airs : set
44:
43:
                            if(atomCounts.ger/atom; =null)itomCounts.put(atom, 1);
                            else atomCounts.put(atom, atomCounts.get(arom(+));
46:
4 - ;
45:
40:
                    mambanes++;
o⁻:
51:
10:
               mr. plage (1)
53:
24.
               // Want sorted list by the atoms' counts.
5
56:
               Set<integer> counts | new TreeSet<intoger>(atom@ounts.values());
77:
                    for (Integén : : doints'
                       for (String val : atom/ounts.keySet())
. 5:
                            if the Thustrager values was a
e 3:
                               System.our.pr ntln.: ": " + val.;
1.1
+ 2:
              catch Exhaprian edg
```

pipeline/Trainer.java

```
: package pipeline;
 z: import | availed.81.8;
 3: import (valid.PriodumputAtream;
 4: import ava....FileReader;
 o: import (ava.in.PrintStream;
 8: import sonemagenerator.CumulativeClassifier;
 r: import weka. Classifiers.Classifier;
 'U: import weka. plassifiers.bayes.Naivedaves;
11: import weka. Plassifiers.rules.Tebusjoniapie;
19: import weka.dure.imstances;
^ ::
^4: **
in: * Takes an arif file to train a decision tree with, and an instance
16: * to classify, initially in DRS form. Produces the number of marks
17: * this instance has been awarded.
19: * Wauthor Jamie
19: *
20: */
21: public class Trainer
22: 3
       public static void train(String trainingFile, String trainedModel,
23:
            String learningType, PrintStream ps)
24:
25:
26:
            try
27:
               ps.printin("** No trained model found. Creating new one.");
28:
29:
30:
               FileReader reader new FileReader(trainingFile);
31:
               Instances instances - new Instances(reader);
32:
               instances.setClassIndex(instances.numAttr(butes()-1);
33:
               Classifier dt - new NaiveBayes(); // default
34 •
35:
               if(learningType.charAt(0)--'d')dt = new DecisionTable();
36:
               if(learningType.charAt(0) == 'j'){
37:
                    // Learning type is either '!' (a standard mark scheme)
38:
                    // or of the form '7-k', where k is the total possible marks
39:
                    // obtainable before it is capped the maximum mark.
                    // (e.g. "Any of the k following, maximum n")
40:
41:
42:
                    We also need to see if there is a 'main.marklist' file
                    .. in the same folder as the traininghile. If so, use it.
43:
44:
                    String markbistFile = getFolder(trainingFile)+
45:
                            "/main.marklist";
                    46:
4 . :
44:
                    if(mark!:stF: e!=null
                       ps.printim:"Using the manually created mark scheme.";
49:
50:
5:
                    if(!earningType.length! = 1 dt = new OutulativeClassilier/
                       darkilistFi e,
52:
53:
                       changeExtension(trainingFile, "markscheme"));
54:
                    else on = new CimulativeClassifler
55:
                       mark..atFile.
56:
                       changeaxtens on (trainingFile, "markscheme" ,
57:
                       new Integer(learningType.spl.t("-")[1]/);
58:
               if (learning Type, charA=(1) 'n' dt = new NatvoPay=s();
59:
10:
11:
               dt.bmliflassifier(instances);
6/:
```

```
ps.printin("############################
               DS.Drint n("EXECUTING TRAINING");
84:
               ps.println("Using learning type: "'rearn waType';
35:
               us.println("** Saving trained model to '"*fininedModel,"'.");
óć:
₹7:
               // Write trained model to file
.01
               OptgatOatbutStream gos = new OptgatOatt itSiteadal
: ۲۵
70:
                     new FileOutputStream (trainedMode) (;
               cos.writeOprost(dt);
, · :
7/:
               oos.flusn();
               pss.dlase();
73:
 4
 r÷:
              catch Exception e [
76:
               a.mrintStackTrace+1:
1 :
18:
79:
80:
        * Changes the extension of a filename.
₹1:
82:
       private static String change Extension (String file Name, String new Ext)
83:
34:
            int pos = fileName.lastIndexOf(".");
85:
            return fileName.substring(0,pos) + "." + newExt;
86:
87:
88:
        private static String getFolder(String fileName)
89:
90:
            int pos = fileName.lastIndexOf("\\");
91:
            pos = Math.max(pos, fileName.indexOf("/"));
92:
            return fileName.substring(0,pos);
93:
94:
95:
        static final String trainingFile = "./files/testing/5aii-0.arff";
96:
        static final String trainedModel = "./files/trained/5aii-0.model";
97:
       public static void main(String[] args) {
98:
            train(trainingFile, trainedModel, "d", System.out);
99:
100:
101:
```

```
1: package cipeline;
3: import wava.id.BufteredReader;
 4: import [ava.1..Fl.eBoader;
 5: import mava.ic.ImparStreamReader;
e: import [ava._ EnthoWiller;
7: import java.net..RL;
a: import java.net.TRLEnmoder;
12: /**
11: * An alternative way of parsing English sentences
11: * which uses an online version of the CAndC parser
13: ' and boxer. This is acceptable for single parses,
14: * but it would be considered antisocial behaviour
:5: * to parse and entire data set.
'f: * @aythor Jaxie
1/: *
16: 1/
19: public class WebsiteParser
20: 7
21:
      // URL to use.
       final static String unl = "http://svn.ask.it.usyd.edu.au/"+
27:
            "demo/demo3.cgi?sentence=";
23:
       final static String url2 = "&printer=prolog&model=a&resolve=yes"+
24:
            "&submit=Parse+and+Box%21";
25:
26:
27:
       // Used in the main method for parsing an entire file.
       static String imputfule = "./files/8b-answers.csv";
28:
       static String outputFile = "./files/8b-answers-processed.csv";
29:
       final static int exampleLimit = 1000; // do all of them
35:
       final static long delay 20000; // 20seconds
31:
32:
       public static String getDrsFromWob(String sentence)
33:
34:
35:
            try
35:
37 -
               String fullUt! = url+URLEncoder.encode(scatemor) (uri2;
               System.com.organin("Reading from: "-fullUrl);
38:
               URF, webbrl = new URL(fullUrl);
39:
               BufferedReader ex[Reader = new BufferedReader(
47:
11:
                    new [nputStreamReager(wepUrl.openStream()));
.2:
               int | 0;
43:
               Strain drsStrand = "";
44:
               while (ur!Reader.ready() && <= 13)
45:
461
47:
                    String line = urlReader.readDine();
48:
                    if(i = 13)
49:
50:
                       int lastBracketFos = line.lastIndexOf(')');
                       drsString = line.slostring(0,last&racketFos,.tria();
51:
52:
53:
54:
55:
56:
               urikmader.cumse();
57:
               if or satisfing equals """) throw new Error ("Problem obtaining "+
τ5:
                    "DRS string.";
5):
               Syst a. . . . . rantin'"Obtained the DRS: "+irsStrang);
# T:
U. I
               refurn disser hij
^
```

```
63:
               catch ("xception c, ;
£4:
                e.ms:htStackTrace();
ES:
                return null;
<6:
. :
58:
0.11
 70:
7 :
         * @deprecated
 /2:
         * (Use of the online parser for an entire data set
 /3:
 74:
         * is strongly discouraged. Use FileParser instead.
 .5:
 16:
        public static void main(String[] args) {
 17:
78:
             try
79:
80:
                BufferedReader br = new BufferedReader(new FileReader(inputFile));
âi:
                PrintWriter pw = new PrintWriter(outputFi e);
82:
                int exampleCount = 0;
03:
84:
                while(pr.ready()&&exampleCount<exampleLimit)</pre>
85:
86:
                     String line = br.teadLine();
87:
88:
                     int textPos = line.indexOf("\t")+1;
89:
                     String score line.substring(),textPos-1);
90:
91:
                     String drsString = getDrsFromWeb(line.substring(tex(Pos));
92:
                     System.out.println("Obtained DRS: "+drsString);
93:
                     pw.pr.ntln(score-"\t"+arsStrang);
94:
95:
                     exampleCount++;
96:
٠
                     System.out.printlr("** Sleeping for 20 seconds");
98:
                     Thread.sleep(delay);
99:
100:
`\.':
                br.close():
102:
                pw.close();
103:
104:
             } catch(Exception e){
- ,5:
                e.bi.htStackTrace();
106:
" ( :
108: 1
```

gui/EvaluatorPanel.java

```
: package _;
                                                                                                                                                                                                                                    To be favored to the character of the edge and
                                                                                                                                                                                   · :
                                                                                                                                                                                                                             light Family in the "d" (
 : import -- -: -: -: -:
                                                                                                                                                                                                                             "n"
                          -: import
                                                                                                                                                                                                                             F: import wish subject;
 Government of the "Decision Tree",
                                                                                                                                                                                      .
                                                                                                                                                                                                                                             "Naive Bayesian Network", "Cumulative Classifier" :
 import what was a manual Apr will rest to
                                                                                                                                                                                                                             is any figure or new order Box reported them togate;
 in in include water = 0.11 hard ;
 9: import (undured) .Fr (e tuling solette if
                                                                                                                                                                                                                             as the condition was about the calculations
  CANAL STREET TENTE SECTION STREET
 Take import (1984-1994) The Take (1987)
                                                                                                                                                                                                                            re decless a new or a residence
                                                                                                                                                                                                                            Time the Limit Products Assess Bake, Live term (2001) 1987 Aug.
 : import the second second
                                                                                                                                                                                     . .
 of import wax.swill. The r
  : import = 19.5% (). Third () 35 of
Tet import avaktewith. The laster
                                                                                                                                                                                                                             Elines its about a new Tolled its or "Status" :
                                                                                                                                                                                   ^ :
 F: import = e.c.x., sw/c st ***ex**A) =55
                                                                                                                                                                                                                             JEarch Star W. V.A. new Tearches,
                                                                                                                                                                                                                             enaturiamed.cen (departmentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentariamentari
2.: import savik.ewi v. Pair Walk ()
21: import syax, swise, borde .Bot .cr;
                                                                                                                                                                                    2.11
                                                                                                                                                                                    64.5
                                                                                                                                                                                                                             er al deAcoa | new lifextAlleact/Lab it
and chânge. From Millions P. 2 T. S.C. ;
                                                                                                                                                                                    . . .
                                                                                                                                                                                   81:
                                                                                                                                                                                                                             Jacroliface son libone new damiel, aneistatusAreas:
endousA mediaeth cout 16 false :
                                                                                                                                                                                    - :
2 . .
                                                                                                                                                                                                                             shightslatel.adoison(lloage);
285 × 4
27: * The pank on which the user for the fitting all will be
                                                                                                                                                                                    , .:
                                                                                                                                                                                                                             of act pattralage: new leage ; ;
28: * of the system.
                                                                                                                                                                                   5.0
                                                                                                                                                                                                                             potterPanel.add res Fo deri;
 29: * Sauthor Lan e.
                                                                                                                                                                                     ٠٧:
35: *
                                                                                                                                                                                                                             addite; canel, Bonderhayout. WRIP.;
55: *
32: public class EvaluationPane extends Flame implements PropertyTherae tall her
                                                                                                                                                                                    15.1
                                                                                                                                                                                                                             Shirt shiptisher , Posterson your . Dangle ;
                                                                                                                                                                                    45
                                                                                                                                                                                                                             as (bartomerase), Borderhayott Schiff;;
 1 . . .
                                                                                                                                                                                    - 4 🛫
                        final 3": mg TASTETT, AT = "Create a new folder, containing either "
 14:
                                         "of the following: \n" |
                                                                                                                                                                                    14 / Z
 J: :
                                         "* answers.csv - a file of the source answers in English.\n"
                                                                                                                                                                                    · i ·
                                                                                                                                                                                                    public void problem to three orby handoEvent on () (
 20.0
                                         "* answers-processed.csv - a file of the parsed sentences.\n".
                                                                                                                                                                                    41.1.2
                                                                                                                                                                                                            if "progress" -= -... > tar tell year . Tar tell
                                                                                                                                                                                 1000
                                                                                                                                                                                                                     int progress = (integer) ovt.ge Nowybluett;
                                         " This skips the parsing stage.";
2 1 2
                                                                                                                                                                                 · . :
                                                                                                                                                                                                                    ing transportant and value in grassi;
3 1:
                                                                                                                                                                                 102:
4.:

    Osed to hap lesse of type leverisons

                                                                                                                                                                                 1 12
4
                         in to their representation we be the co-
                                                                                                                                                                                 100
                        final Magk2 tego: ,%i ingo men Map = new dashMopelonseger, Strong>) ;
 г:
                                                                                                                                                                                                            public class Evaluates/ifico. implements Acta Elisteral
- :
                        final Cextoneu sist date of
                                                                                                                                                                                     . :
44:
                        is the char e of learning tyrus.
                                                                                                                                                                                 1 7 5
                                                                                                                                                                                                                             public void and embedianced Admical ventues
                        In the Kill-ord old term
- :
                                                                                                                                                                                     . ·
                        Chicare assi promes .ar:
 d):

    Frush textarea

                         To the section of the
 . :
                        final integrada was restorior of new rootSituate of this,
                                                                                                                                                                                                                                              BINKSUFASCHIECTINAT "" ;
                                                                                                                                                                                       :
                                         "Test directory", "DIRECTORY";
 . ':
                                                                                                                                                                                                                                                   Ger leaktina tirk
 5 :
                                                                                                                                                                                      _ :
                                                                                                                                                                                      2:
                                                                                                                                                                                                                                              final first adding the applies were
                        public : tal rate: " The U.)
                                                                                                                                                                                      1:
                                                                                                                                                                                                                                                              learner illation selvete describer
 - . :
                                                                                                                                                                                      - :
 5.3.2
                                                                                                                                                                                                                                              final intrinction to the Mand of the thing in the termination of the final control of the fin
 - .:
                                         Barness, in new Districts and it is
                                                                                                                                                                                                                                              if Theory Dona - null states a green rate
-- ):
                                                                                                                                                                                                                                                              "ERROR: No folder selected" ;
                                         性性性的 regression new 框面 someworksome ay (Fig. 1) 1);
 -. :
                                         lea nî danî en dian î e new lêsh lo :
                                                                                                                                                                                                                                              else
                                                                                                                                                                                                                                              at with in history as as false;
                                          or trextAres instructions now Stex. Breatt, 507;
 - 21 -
                                             COST NUMBER OF STA
                                                                                                                                                                                                                                             le this into the bugoithment refereds broken
                                                                                                                                                                                                                                                          on it is also and figure in a figure
                                         ating Bulleton new date to "Start";
                                                                                                                                                                                                                                             Tack task new ask follows with, lesson styre ;
                                         at right to this term of the new oralist into the contract.
                                                                                                                                                                                    . . . .
. .
                                                                                                                                                                                                                                              mastril terminagi en en si am di this ;
                                         Line burn disconnected that Follow to
0.25
```

gui/EvaluatorPanel.java

```
1.23.44.4.1.18. ()
 + :
  :
             If the contributions that the base of contribution of,
            public class area extends Formount work floor, first
  :
 · - :
                     r - a Transmir r- r
                    public like String to booten , Comercial programs
                           this.: 'frometare = frometare:
                           this.learnangTv: learnangTv:
 . :
 . :
 r' :
                    public ...a aconFarkor ind
 ٠-:
 ٠.:
                           if learningType.og.als "j"
 r. :
 d -, :
 4
                                   String s
١٠, :
                                          (Stilling) Soptions and all without the
15:
                                          "OPTIONAL: If the mark scheme is of"-
152:
                                          " the form 'Any of the k following,"-
153:
                                          " maximum n', enter k.";
151:
                                   if(s!=""&&s!=null*learnameType "j="..;
155:
156:
                           Tester.test(folderName, learningType,
1500
3 - 61:
                                  new GUIOurpur(starusArea), or sneed of;
5.23
                           return null;
11.15
1.1:
                    public void dose() (
10.72
1.00 5.2
                          PolikinigenDebalinToolkit().heep();
1000
                           stairBurrom.eetEnabled(true ;
. . . .
                          reservingson mull ; which off the wait his in
. . . :
. :
```

gui/FileDragPanel.java

```
: package .:
 -: import statement it
 : import - -: .....
 : import fire.so . 1 % 3.5; .;
                                                                                      ٠:
 is import to late. Identify at its 12 to 12th 17th
 is import to a swind decision in the line of said ()
 : import first. and . T. T. T. T. T.
 : import | a alay lamble | a Partier;
if import (1919) in the late of the first the LA (1919)
                                                                                      ·.:
4: import (20.380,21 ) | 1 10 0 197 (7- 1)
import ....av ...all y a set at= ;
                                                                                      . .
"v: import analysts wert. Modern with
                                                                                     ₹ /:
17: import 1000.0W .=.0011.V .:- 120001019
0.13
                                                                                     -1:
all import which we had been to
20 : import | winds.combacinese Incom-
                                                                                     - 3:
22: import savak.swing. in an out ;
                                                                                     ٥:٠
27: import (2004) 1.7:1009;
                                                                                     85:
                                                                                     ēs:
24: import svax.swarp.draper;
                                                                                     87:
ző: import avax.2025a, az ij
Wet import havex. swine, ford r. bitube;
                                                                                     ár:
                                                                                     89:
27: import -avyx. swims.c room.l distribute
2h: import graxiswamach garg.CombaBondam;
                                                                                    9J:
29: import avax.awinq.triedmo ser.F leftitel;
                                                                                     91.
                                                                                     92:
                                                                                     771
vuit in Alaghare pamel char represents à 1916 or forestail
                                                                                     07:
                                                                                     951
Fig. 4 sole tight. The upple can specify a Tile by up of
44: * Two methoda:
                                                                                     26.3
                                                                                     37:
and it is a discretify the parent usa indentione have recently.
Pre to the normal fashion.
                                                                                     250
ent of the dead a filterenie toer fund the box.
                                                                                     Cala:
العرابيل أرامعيها الالترات
3 : *
                                                                                   11.7:
41: public class Fit (Printing) extends (Finel implements Drop argoth state)
                                                                                     . . .
                                                                                    1.4:
4
2 . .
           static [mainstants] = "./icons/document.gif";
                                                                                   0.5
           static annual and Finally yet = "./icons/document_greyed.gif";
                                                                                    · . .
শৰ:
- :
7. :
           final static Color Host Gulebild's new of the color says of the
                                                                                    14:
                                                                                    1.41
9::
           28 gTurn = .cr
                                                                                   . . :
10:
             Thamel for Himphace ;
                                                                                   11 :
             Committee of the state of the state of
             Corton filelyse;
. :
             37:12:3 75:37:
             11.ce/ 3 [11.16]
- 12
                                                                                    · . :
8.12
             Figure vol.;
                                                                                  1 / 1
             2 a Home and High
 -:
             public Programme = New return : Learn :
. . .
             public to define and to the boundary of the product of the
. 2
                                                                                    . . :
                                                                                    1. :
1 1
               super ;
                                                                                     16:
               this:
```

```
Fig. 2000 new man 1 of the pro-
        forether and settlement of the contract of
       Fig. '- new Fig. "Tahoma", _ '. . . . . ;
       For experience to the second of
     Time new Title new to the partition;
      in Those new closed new coast Table (% 1) as well as
        The second of th
      The sale of dates, in the more thank the p
     months in the content of the first section of
      Postario to new international new content of the content of the property of the content of the c
       231 3 3 7
     Form 15712 - new room "Tahoma", P. Mill Alic, W. J.
     in LaType worldware in it to gr
               definitions toxy a pairwise ease mogranic
                THE Case will is made from events
     u = new ObligTirg(f)this, this;;
public void drag magn trug Temper Temper on Se, . . .
public void drantxi;(T) or arast vent d+e' { ]
public void dragover (Crop large transfer no diste, in
public word Stateon of Tie mentals apprehing apprehing the mentals and the mentals are presented as the public word.
public void that it planuable up them above
            // Get the dropped object and try to figure out what it is
            To renewate on a surface with 1 table beings on the
            Callablay it if wors from the Transfer Natia Flavorsity
            for int = 1; _ < fuer representation =
                                Check for tile thats specifically
                        if ((lavora))).isFlavordovacielistTypo ,
                                                          Timari Modani noon dhoo t...
                                                 attle.sectorop(DetConstants.ATTINE 07Y_00_M Ver;
                                                      The contract "Subtraction is e , then eta_2 \circ e_1 , e ar{e}_2 \circ e_2 , e ar{e}_2 \circ e_3 , e
                                                  Tavalicit.First ter
                                                        iname to Turset transcribes stephenu bloom galib :
                                                 AttributePanel missificationetry, p. ......, ;
                                                 1 It we make it thes for, everything worked.
                                                  at Make tilliplets true ;
                                                 return:
                                     Ry hallstell dustrant To a to get 2.
                       else if flav reliation of the transfer all adoptions are
                                                The sections "Since of Contest of Minds of the
                                                  / Gosect D = ir, retiralist=rP (La, zi no-s, 1, );
                                                        Turk against a " a linear "
                                                 disease a faithre true';
                                                 return:
                                 Hory about an inner streeth.
```

7. :

gui/FileDragPanel.iava

:

:

```
else if the control of the control o
                                                                                                                                                                                                        1 1-1-27 12 13.
            - :
                                                                                                                                                                                         1 : . : _ _ _ true ;
                                                                                                                                                                                    return:
               :
              :
                                                                                                                                          of the control of th
              :
                                                                                                             + 2
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         . . :
                                                                                                                                             catch
       57.2
                                                                                                                                                     إربوديا الأي ييابين
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        1000
        T ( : )
                                                                                                                                                                               the part of the pa
     ٠;
     ٠:
                                                                                        public void attraction of the spit earth
      · :
                                                                                                                                             this. The management
      --:
      4 . .
                                                                                                                                             TO NOTE HE OF STATE ON STRUCKS FOR THE STATE OF
                                                                                                                                                   . - -
      a · 1
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         2.55
                                                                                       private static in a symmethic gard per whole only to
      · :
   . ::
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        . :
    · ᠳ:
                                                                                                                                       int the Markinex collegement astondex's'"\\" ,
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        2.73
    500
                                                                                                                                                              t.leftape.lasrIndexnf("/";+;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        / ":
                                                                                                                                       3: nu li =Name(n.ly = illeName.substitungiquar =);
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        203:
    -1:
   F/:
                                                                                                                                    if the left mesonly, length(,<=17 return fileN month by;</pre>
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        214:
                                                                                                                                       / Otherwise we need to shorten it.
    53:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        215:
    54:
                                                                                                                                       - 2 ROSAL Find dat
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        2 6:
   55:
                                                                                                                                     int p s of = forcharsOn y.lastIndexOf(".":;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        2 7:
1551

    Penning andulas.

                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        ž 8:
1573
                                                                                                                                       return ' of askirly substructed, O+",,,"-
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        213:
                                                                                                                                                           tic (Name Only, sugarning, wasPane or;
      101
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       24:
       • :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        22 :
  . :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        // :
  · :
                                                                                      public class : : : : implements Volselistics
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         2233
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         Dun:
  . :
                                                                                                                                        TB:1 ... in or new TB:140: / ser ;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       7 Ed
   . .:
                                                                                                                                        final instruction ;
                                                                                                                                       public of the final fit was taken
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        ٠- :
                                                                                                                                                                                    this.Typ: - Type;
                                                                                                                                                                                       \tau_{\perp} = \tau_{\perp} = \text{new } \forall yet c : (1 - 1.6 - )
 6.55.2
                                                                                                                                                                                    if Tyre, qualse "DIRECTORY" (Fig. 10) Selection Mages (
                                                                                                                                                                                                            THE Offices INTRECTORIES WAY :
                                                                                                                                                                                   - reservational servicing
                                                                                                                                      1
"::
1743
                                                                                                                                       public void on sanit oxid Mossefvort
                                                                                                                                                                                         Hardle open barron action.
                                                                                                                                                 int funda. Shishimitatisalim terk t
       ٠:
        :
                                                                                                                                                  if the many at the TRibe at user Large Fig. 1.1.
                                                                                                                                                            Double of Asterbeland Maile :
                                                                                                                                                                         100 miles | 100 mi
        ;
          :
                                                                                                                                                                                       Control to the state of the second of the second
```

```
biov sildua
                   1 27 F
                   public void
      public void
                          .....
      public void
                   ~
public class MFLLE ... extends
      Friend i elver:
      public My it is a district the same
            this, the view of the Pype;
       public at a talk the training
         return fig. . . . . "DIRECTORY" . "Directory" :
            0.0.2.5 -0.02.
       public boolean arms: .
                Tan mage 1 42 group of 1999 to 12 ero elektronia.
                we have the common deserting as with the
              if til-.lik to the return true;
              in the when a maximum filler.
              if excess (! null
                    if exper, bull-made fore ype :return true;
                     else return false;
              return false;
       public String getExample on the conf. [
               Str bull;
            Storogist Community (1)
            int = p.lasting backing.
            if the section is the
             two sign of the in line lase ;
            return -x::
```

gui/GUIOutput.java

```
: package ; ;
                   : import - ... - - - : : ;
                      1 import | 1 28. / 19.2278 & ent
                 The state of the s
                        Tokon temperatur
                                                                                                                                                                  public French Plank Arta
                                                                                                                                                                    super y. e. . this...
         . :
                                                                                                                                                                  public void in the first that the support of the su
    . H.:
                                                                                                                                                               public void promise transpar
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              through the first
            • :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             (-1)^{2} \cdot (-1)^{2} 
                                                                                                                                                                    public void promount to
  21.5
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    talappen: "\n" ;
                                                                                                                                                                       public void graph land
_ I
         2: 3
```

gui/MarkerGUI.java

```
: package ;
: import
           :
. . . . .
 : public class '... : I
          public static void
          .: '- ....' new Til - "Marker" :
. .
               ...
             try
 :
                \text{MIM}_{\text{cl.},\text{total}} = \kappa N \cdot \sqrt{2} \log t.
. :
                  SIManager.gelfystema.ckArd.com/?rawa4/nengy);
- :
...:
               catch 'Expertion ex 100
. ::
24:
21.:
             Proceediane tales new MacbedPane();
2 ...
             factions rate Pate "Pipeline",
/T:
              new Image cor("./icons/palmpilot.gif"),
76:
               new !upc! reranel(/);
24:
             ricos.a ea Palat "Evaluator",
              new Image For ("./icons/cog_small.gif"),
5.10
7 F:
                 new EvaluatorPanel());
             iscolabel"Marker",
:7:
3.52
               new Imageleon("./icons/construct.gif"),
                new MarkemParolly);
             time.set unemplane/ .add:tabs()
...:
115.1
             than indicates, p
. . .
٠:
```

gui/MarkerPanel.java

```
of a section of any area of the first pro-
       ; package
                                                                                                                                                                                                                                                                                                                                                                                                                      . :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              : import - " . '.= i- . ';
       : import : : : : :
      : import : ;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             Annual Color Data Engineering to the Annual Color
      import that while taking it;
   import property contains a contain a cont
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              This is a series of new Notice new or forming the
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               included new at- "";
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  new -- "Tahoma", - . ! C , P ;
       I also to stickled to leave and in larger of $1
    Contract to a visit of a substitution of the
for import of all the burns
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            to the sinew order of the state "Mark";
  .
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            and the second of the control of the
19: import 30:3, 21: .5987
                                                                                                                                                                                                                                                                                                                                                                                                                 80.1
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             in a share issenting or all all
                                                                                                                                                                                                                                                                                                                                                                                                                15 :
The import avak. With a company
y : import atax.aw top. The = ;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      where the case of the top top project best on.
public class Cath. "Actaunicanage implements A chang prores
.1: import | Diskiew hp. 1.4x7 i 1:4;
                                                                                                                                                                                                                                                                                                                                                                                                                ~ < :
The import of the section of the section
                                                                                                                                                                                                                                                                                                                                                                                                                 - - -
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      public void intichternommed(Assismiveme c
14: import of axisy trings length damages;
                                                                                                                                                                                                                                                                                                                                                                                                                 5 15

    Arif and model files must be stop i ed.

                                                                                                                                                                                                                                                                                                                                                                                                                 58:
A : import ransen.OPSreador:
                                                                                                                                                                                                                                                                                                                                                                                                                89:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             if (note: a.get) to -loars = mull return;
21: import probleme.WapsitcEntage:
                                                                                                                                                                                                                                                                                                                                                                                                                500.5
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            if mount, File manufile Ware -null return;
The import weks.c.ass I e avillastic if
                                                                                                                                                                                                                                                                                                                                                                                                                449 :
 dat import wekali nel marabaka
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         Strain to Massity = Indefance .or ross Scalence.
 and import weeks. Tile. In let fire /
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             abawa:.pa(Text/));
51:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             If runninger (greater = WegenroEasser.me(DraintomWeb)
                                                                                                                                                                                                                                                                                                                                                                                                                   _ .
 run it is pursuable to serve a minimum alone by one and
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              At SWell PretText() ()
                                                                                                                                                                                                                                                                                                                                                                                                                45.
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              SetrStructure set = DRGreader.at m_1(z_0, z_0, z_0) inserty,;
 Page 7 receive in any idual charactication.
                                                                                                                                                                                                                                                                                                                                                                                                                673
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             mist<br/>Kõduduek valis = new maaksi.istriluugilek (j.
 Fr. & Stanish of June 12
                                                                                                                                                                                                                                                                                                                                                                                                               981
· 6: •
.7: */
                                                                                                                                                                                                                                                                                                                                                                                                              44:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             try
 18: public class Werker and extends Tele-
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         Butfal ateracy by new Butt restains
                                                           는 회사교회관 및 함께 한 회 new 한 분들의 athanel this, "Arff", "arff", ;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 new Prilamprograms (not be adopted by Name of the
                                                                                                                                                                                                                                                                                                                                                                                                             1 2 *
 . . .
                                                          is a section for the first new bulb repParolethis, "Model", "model";
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            Thatening thereint (m/"** Reading ARFF.");
                                                         TO AND SELECTION
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         boolean reading true;
                                                           El El El el de sette caracing.
· . :
                                                                                                                                                                                                                                                                                                                                                                                                                    . .
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          int i=:;
47:
                                                                                                                                                                                                                                                                                                                                                                                                                 2.15
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           while to the early take to into
                                                        public Markettarer
45:
                                                                                                                                                                                                                                                                                                                                                                                                           1392
4. :
                                                                                                                                                                                                                                                                                                                                                                                                              1500
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 ransism in the residence of the figure of th
                                                                                          Haribayout, new Element ago in the
4. . .
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 if we are in 1 18811 File vix in "@attribute" viv
400
                                                                                             Fig. 1 to 18 to 1 = new (East of new Pisiensavo) (1)
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              ijadini false;
/<sub>T</sub> *
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 else if limetimpoxiti"@attribute" and A
                                                                                          그 번호 번호(22개호 5/3 ) 13 5 new 오고 하드면,274,25 한 것
 ī :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          - Storegall grazile = line...enice # #.j.
  ÷ :
                                                                                             · Funda 는 10 전환 시 1 · new 전 전 new Pu Wildy 1 it / 수
                                                                                                                                                                                                                                                                                                                                                                                                                  ´¬;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 if there .compare "score", .
                                                                                          າງຕາມເສັດເປັນ ເພື່າ new ພົລຄຸດເຕ"Sentence: " :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       if something the party of the party is a first of the something of the so
   Ī :
                                                                                            now timew Taxite or 7
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             (0.00.03)
   E .:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       else valetita ili ;
                                                                                          impur Furth 1, as a littlewest of
                                                                                             new E :- 'Submit" ;
                                                                                          As the best track to be a single of the mew districts of the interest of
 Ī:
                                                                                          logic of the Borotom Paragraph of the state of the state
                                                                                                                                                                                                                                                                                                                                                                                                                 ¿ :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                - ;
                                                                                          Control of Charles and the
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         literation new literation are "Input";
                                                                                                                                                                                                                                                                                                                                                                                                                  _ :
                                                                                          The Frederic Frederic ;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                Simple Control of the Control of the
```

gui/MarkerPanel.java

```
. :
           . :
                                                                                                                                         int
                                                                                                                                                                                           in the second of the
                                                                                                                                                                                                     - = new double ;
                                                                                                                                         double
                                                                                                                                        -:
           .:
                                                                                                                                        The training of the state of th
        · :
                                                                                                                                                    a substant of more companion with the
        * c :
        5/:
                                                                                                                                            the most of the most of
        575.1
                                                                                                                                           them I Introduction on the selection for a settle contains Care ;
         1115
                                                                                                                                         The state It like East the like of the state of the state
          . . .
                                                                                                                                           2010/01/01
                                                                                                                                         the argument value takes onew line of \kappa = 1/\rho + 1/2 \approx \gamma
       . . :
                                                                                                                                         Colors and the limit of new FilePrices of the Color of the Color
     44:
                                                                                                                                         up for a other errors onew Transit was become a
          Ξ:
                                                                                                                                              a lester largeries all Attrice (
                                                                                                                                        . .... har elsetőeteset instances )
    100
                                                                                                                                       double a management placement by a management of a
ey ''' . . .gringin "Classified as: " ;
    - ` :
                                                                                                                                       of been right rusettextionew union of a constValouts=""";
1-/:
157:
154:
                                                                                                                     } catch() x nother (x){
1555
                                                                                                                                                    ιχ.prinsβhaekmmaecis;
156:
15.7:
15.81
1000
    7: 1
```

: package ::

gui/PipelinePanel.java

and the second second second

```
:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    Fig. 78 (88) (E.S. C. Lawrence of the new control o
    Table 1 - Control of the Control of 
     It amport the way were amount of the
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        : :
        .
     To import that a till garden;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      \exists x : x \in \mathbb{R} \quad \exists x \in \mathbb{R} \quad
    :: import
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     The Court American the industry a new londer content of
    5 E
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     gage anexal and in the State of the group
        : import 100 x.50 12. 11.
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    141-14:50
     in import of Kinabagas group of -;
    v: import = v.x.zwz.j.. = -;
     1 - 1 - 1 - 1 - 1 - new 1 - 1 - ">" ;

    iii) " = 1 km in that the later is new in constitutions on the

        : import wax.swi.bluCtV - i;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    State of the state
     In import waxashing. The contract
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     Paragraph and Char
    c: import parkx. se with this black der;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            A: import page of Card Rooms of
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              -:
     /: import gig= 1 %. 1 = Pits= ;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               っ:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  public class Fig. : Mistomer implements A fig. - ear
 . : import till = ine.'iii = ;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        . :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       public void soli seen simedia for also for
2.5
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        - 3:
 .-: * This pixed allows the used to interior wash
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       · 4:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        Satisfied someonic contents of a second State of Fig. 4 (sec. ).
 2: * - tige of the parsing reason for toposine, and
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         Strane complete administration of the first property.
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      56:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         20: A rowerts between the 4 tyle rethers usual
 green * 1. Bource (SV | Indiath Feather Co.)
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       8-7 <u>1</u>
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          STRUMENT AND TOX "ERROR: Files not specified
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        1 - 11 - 1
  Pri * 2. Bemontie PSV - parseu sentembers.
 Dis * 3. April | 1 classifier apple 1. p.
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    na.
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            return;
 Err 4. Model - A trained one Bulletin
i : * Panthon Jamie
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      TO COMMISSED AND CHAIR PROCESS OF STREET FOR THE LONG TO A THE TRAINING THE TRAINING TO A THE TRAINING THE TRAININ
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        or amor, new Goildoute we exercise teach of
23: 1.
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    : :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      .2:
 in: public class apeliable. Firelain
                                                                           FileDown at a rest of the new FileDown and this, "Source CSV", "csv";
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   public class first publisher or implements Admironthagener.
 38.3
                                                                             To give this to garage to me new Fried to make this, "Semantic CSV", "csv
  1:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      public void ant and refer formed (Anatomi ventor)
                                                                               is equagramed as firmar new title magrames this, "Arff", "arff" ;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         String according Name - Frankin (1800.get) is cone..;
                                                                           it = rapFormal A gae' 139 = new FileDrace and r this,"Model","model";
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        . . . .
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       23.5
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        Similar Composition Name of the Security For Engine of
                                                                                 Tax : A - A s in tight and
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              · · :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         if (some percomber null operate letter enull).
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   . 1 :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           Concentrate of the Control of the security of 
      ŀ:
                                                                            public ? to lanchan+! ...
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         ?2:
                                                                                                                            Herlay or (new rors strongs of m, i )
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             return:
    14:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      . . .
     -:
                                                                                                                            Ronder State of the new in this will "Status";
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     _4:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        william at a new of fur a greatern of;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        Additional and a section and the section of the sec
     . .
                                                                                                                              TERMS Later of the comment of an in-
                                                                                                                                ingur as and coefficients on a service of
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            , false, new introduction and agency
   .::
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               1000
 49:
                                                                                                                               otherwise and new Tracking of synthetic
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                10 :
 • • :
                                                                                                                                다 시민들은 마이크 (Table of new 2 propulate Propulate Pro
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                i c :
                                                                                                                             ica dawn ein tractionia false ;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       - , <u>:</u>
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   public class Born Digiting implements April reparener
    ...
                                                                                                                              in non a electron Dale :
     5:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      Mag state digres new Warsh;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         ::
                                                                                                                              a_{ij} \in \mathcal{A}_{ij} for a_{ij} = a_{ij} \in \mathcal{A}_{ij} and a_{ij} \in \mathcal{A}_{ij}
     . :
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            = i i plate.com "Decision Tree", "d" ;
                                                                                                                                        العداية مخيير لهجير يسريان
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 The interpretable "Naive Bayesian Network", "n" :
                                                                                                                             There is the five new Fire Europe
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          The to Mivier. Dit "Cumulative Classifier", "j";
                                                                                                                            to the to electifized to new Ellowing hit of
       ٠:
                                                                                                                            to the total service of the property of the service of the service
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         public void in the later to be a present
```

```
02/14/08
18:32:31
```

gui/PipelinePancl.java

```
return;
                            Political Continues of the analygical Kind of the Acets
у.;
                            \mathcal{J}^{n} = r(r) . The distribution of each some state \mathcal{J}^{n}
                                              . . . . . . this,
                                              "Please choose a learning type",
                                              "Learning Type".
  ٠.:
                                              null,
guaran _ _ ee,
  . . . .
  1.50
                                              "Naive Bayesian Network";
  37:
                          if(st null) transer. Lra no outgeb TeFame, compute TeNam
e, lest ingreposition : "", new OHTO on this Arealt;
```