Assignment 2 - Using Weka for Text Classification

Given a preprocessed document collection, please conduct document classification using Weka

Dataset:

webkb-train-stemmed.arff

webkb-test-stemmed.arff

WebKB containing 2803 training text data and 1396 test data. This data set contains WWW-pages collected from computer science departments of various universities. These web pages are classified into 4 categories: student, faculty, project, and course. The data set has been preprocessed with removing stop words and stemming. The dataset is already converted into .arff format which can be directly import into Weka.

Method: please pick two classifiers (e.g., naïve bayes, svm, decision tree) in Weka to conduct text classification and return the classification accuracy.

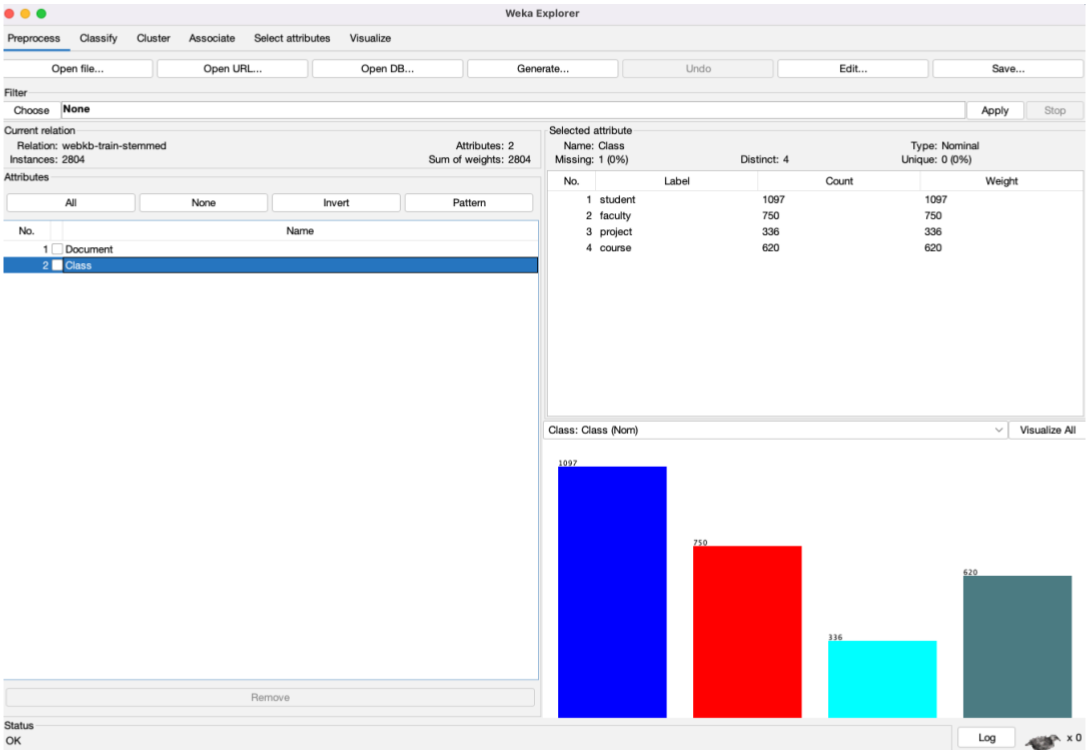
Report: please write a report including the screenshots of generating document-word matrix, loading the given dataset into Weka, conducting classification using naïve bayes. Please specify the parameters you choose if applicable and show the classification accuracy.

Report

I. Loading the training set into Weka

There are 4 classes as the picture below.

There are total of 2804 instances with 2760 distinct values for the document attribute.



Filter: None

Relation: webkb-train-stemmed
Instances: 2804
Attributes: 2
Sum of weights: 2804

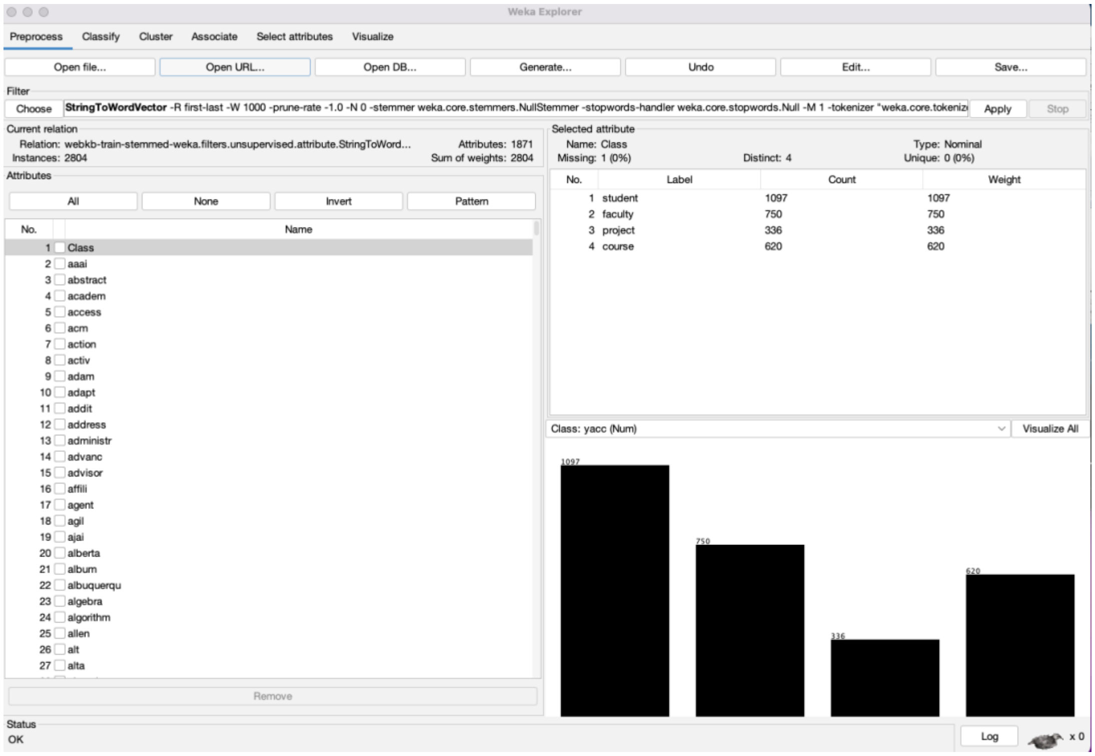
No.		Name
1	□	Document
2	□	Class

Name: Class
Missing: 1 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	student	1097	1097
2	faculty	750	750
3	project	336	336
4	course	620	620

Status
OK

II. Generate document-word matrix into training set



Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter

Choose **StringToWordVector** -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M 1 -tokenizer "weka.core.tokeniz

Apply | Stop

Current relation

Relation: webkb-train-stemmed-weka.filters.unsupervised.attribute.StringToWord... Attributes: 1871
Instances: 2804 Sum of weights: 2804

Attributes

All | None | Invert | Pattern

No.		Name
1	□	Class
2	□	aaai
3	□	abstract
4	□	academ
5	□	access
6	□	acm
7	□	action
8	□	activ
9	□	adam
10	□	adapt
11	□	addit
12	□	address
13	□	administr
14	□	advanc
15	□	advisor
16	□	affili
17	□	agent
18	□	agil
19	□	ajai
20	□	alberta
21	□	album
22	□	albuquerqu
23	□	algebra
24	□	algorithm
25	□	allen
26	□	alt
27	□	alta

Remove

Selected attribute

Name: Class Type: Nominal
Missing: 1 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	student	1097	1097
2	faculty	750	750
3	project	336	336
4	course	620	620

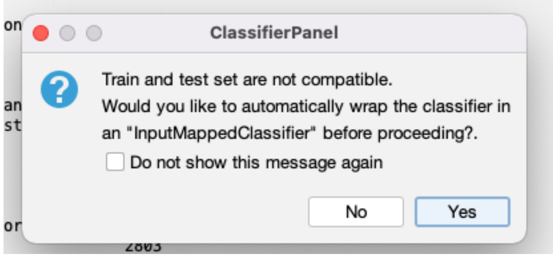
Class: yacc (Num) | Visualize All

Status
OK

Log | x 0

III. Evaluate test set

- If I just upload the test dataset and run the classifier, it will pop up the warning box below. There is a problem evaluating the classifier because the testing document is arff file with string attributes while training document is an arff file with word attributes. That's why they are not compatible.



ClassifierPanel

Train and test set are not compatible.
Would you like to automatically wrap the classifier in an "InputMappedClassifier" before proceeding?.

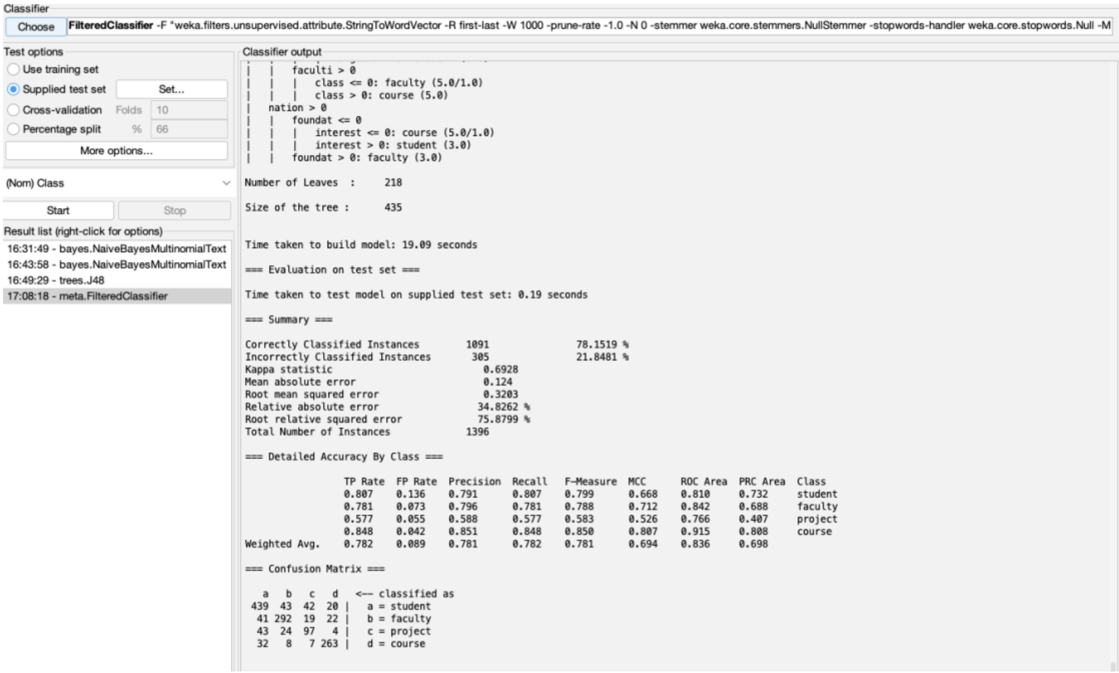
□ Do not show this message again

No | Yes

- Since we cannot just apply the StringToWordVector to the testing dataset, so we can use FilteredClassifier that will create a filter from the training set and use it for the testing set.
- I will undo the StringToWordVector for training set so that it will be the same as Figure 1 above. Then I go to Classify to choose FilteredClassifier and apply filter of StringToWordVector with the following classifiers.
- Below is the list of classification method that I applied on the testing set in the order of ascending accuracy

1. J48

Accuracy: 78.1519% which is lowest. ROC Area is 0.836.
We can see that J48 is not a suitable learning scheme to use on text data.



FilteredClassifier -F "weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Class

Result list (right-click for options)

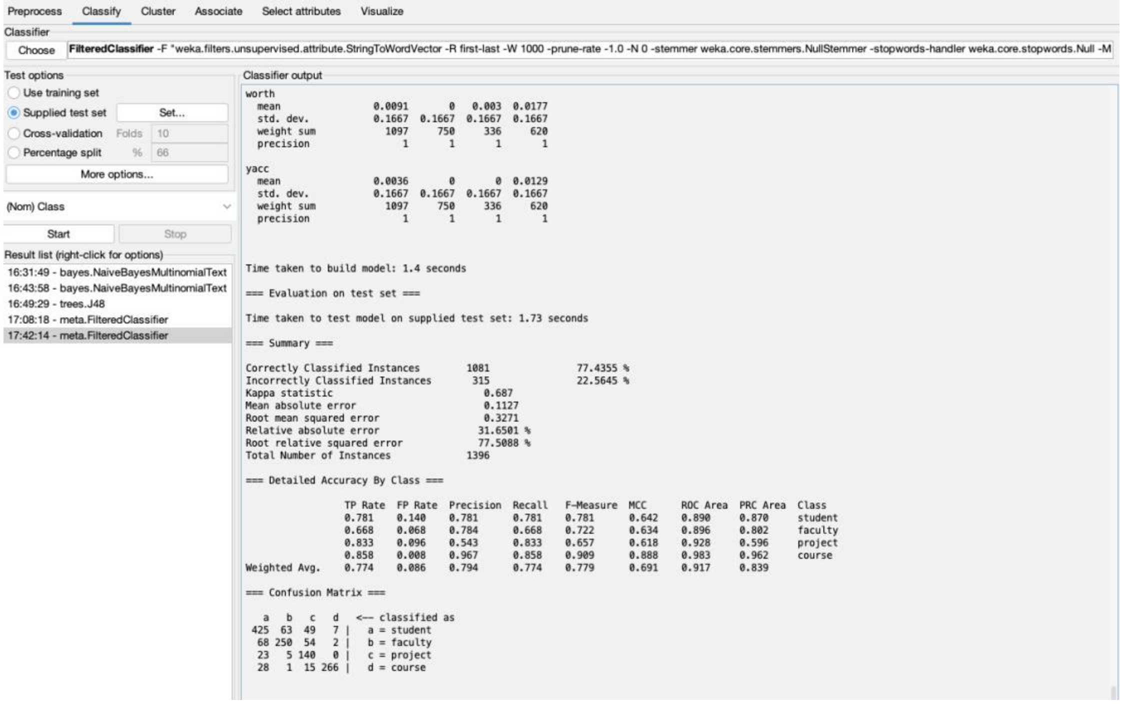
16:31:49 - bayes.NaiveBayesMultinomialText
16:43:58 - bayes.NaiveBayesMultinomialText
16:49:29 - trees.J48
17:08:18 - meta.FilteredClassifier

```
|   |   faculti > 0
|   |   |   class <= 0: faculty (5.0/1.0)
|   |   |   class > 0: course (5.0)
|   nation > 0
|   |   foundat <= 0
|   |   |   interest <= 0: course (5.0/1.0)
|   |   |   interest > 0: student (3.0)
|   |   foundat > 0: faculty (3.0)

Number of Leaves  :     218

Size of the tree :      435


Time taken to build model: 19.09 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.19 seconds

=== Summary ===

Correctly Classified Instances        1091               78.1519 %
Incorrectly Classified Instances       305               21.8481 %
Kappa statistic                          0.6928
Mean absolute error                      0.124
Root mean squared error                  0.3203
Relative absolute error                 34.8262 %
Root relative squared error             75.8799 %
Total Number of Instances             1396

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.807    0.136    0.791      0.807    0.799      0.668    0.810     0.732     student
                 0.781    0.073    0.796      0.781    0.788      0.712    0.842     0.688     faculty
                 0.577    0.055    0.588      0.577    0.583      0.526    0.766     0.407     project
                 0.848    0.042    0.851      0.848    0.850      0.807    0.915     0.808     course
Weighted Avg.    0.782    0.089    0.781      0.782    0.781      0.694    0.836     0.698

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 439  43  42  20 |   a = student
  41 292  19  22 |   b = faculty
  43  24  97   4 |   c = project
  32   8   7 263 |   d = course
```

2. Naïve Bayes

Accuracy: 77.4355% which is low. ROC Area is 0.917.

Naïve Bayes is good when it comes to independent attributes (independence assumption). However, Naïve Bayes treats all words the same, accounts the multiple repetitions of a word, and counts non-appearance of a words as strong as appearance.



FilteredClassifier -F "weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Class

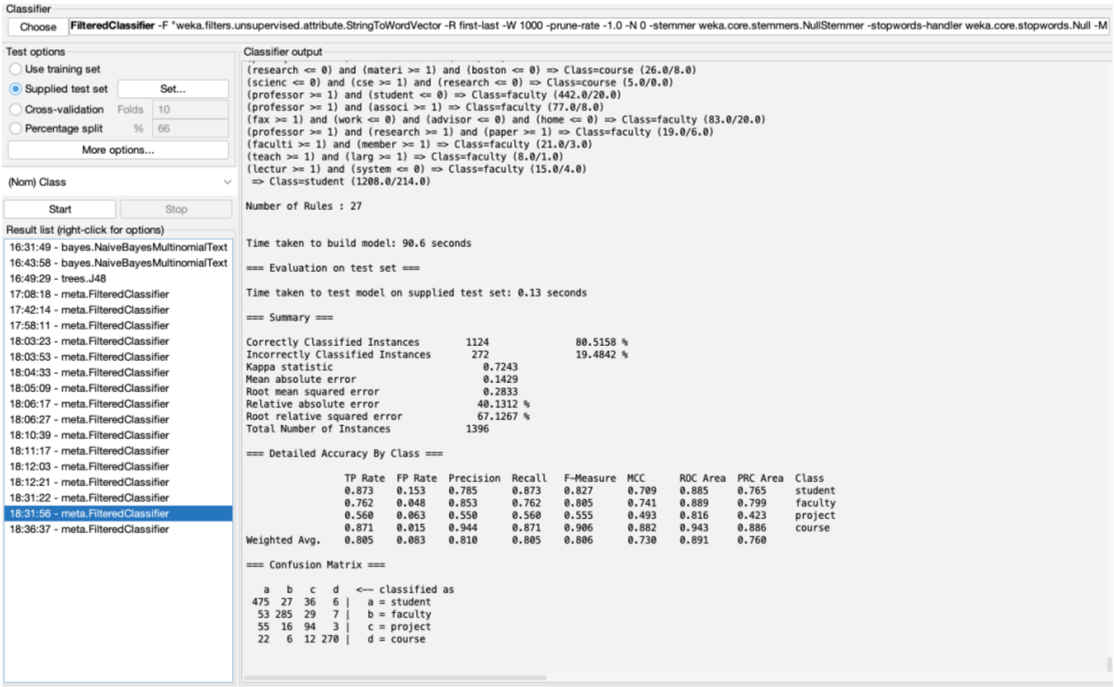
Result list (right-click for options)

16:31:49 - bayes.NaiveBayesMultinomialText
16:43:58 - bayes.NaiveBayesMultinomialText
16:49:29 - trees.J48
17:08:18 - meta.FilteredClassifier
17:42:14 - meta.FilteredClassifier

```
worth
  mean                  0.0091      0   0.003  0.0177
  std. dev.             0.1667 0.1667  0.1667  0.1667
  weight sum              1097    750     336     620
  precision                  1      1       1       1

yacc
  mean                  0.0036      0       0  0.0129
  std. dev.             0.1667 0.1667  0.1667  0.1667
  weight sum              1097    750     336     620
  precision                  1      1       1       1



Time taken to build model: 1.4 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.73 seconds

=== Summary ===

Correctly Classified Instances        1081               77.4355 %
Incorrectly Classified Instances       315               22.5645 %
Kappa statistic                          0.687
Mean absolute error                      0.1127
Root mean squared error                  0.3271
Relative absolute error                 31.6501 %
Root relative squared error             77.5088 %
Total Number of Instances             1396

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.781    0.140    0.781      0.781    0.781      0.642    0.890     0.870     student
                 0.668    0.068    0.784      0.668    0.722      0.634    0.896     0.802     faculty
                 0.833    0.096    0.543      0.833    0.657      0.618    0.928     0.596     project
                 0.858    0.008    0.967      0.858    0.909      0.888    0.983     0.962     course
Weighted Avg.    0.774    0.086    0.794      0.774    0.779      0.691    0.917     0.839

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 425  63  49   7 |   a = student
  68 250  54   2 |   b = faculty
  23   5 140   0 |   c = project
  28   1  15 266 |   d = course
```

3. JRIP

Accuracy: 80.5158%. ROC Area is 0.891.



Classifier

Choose **FilteredClassifier** -F "weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M

Test options

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Class

Result list (right-click for options)

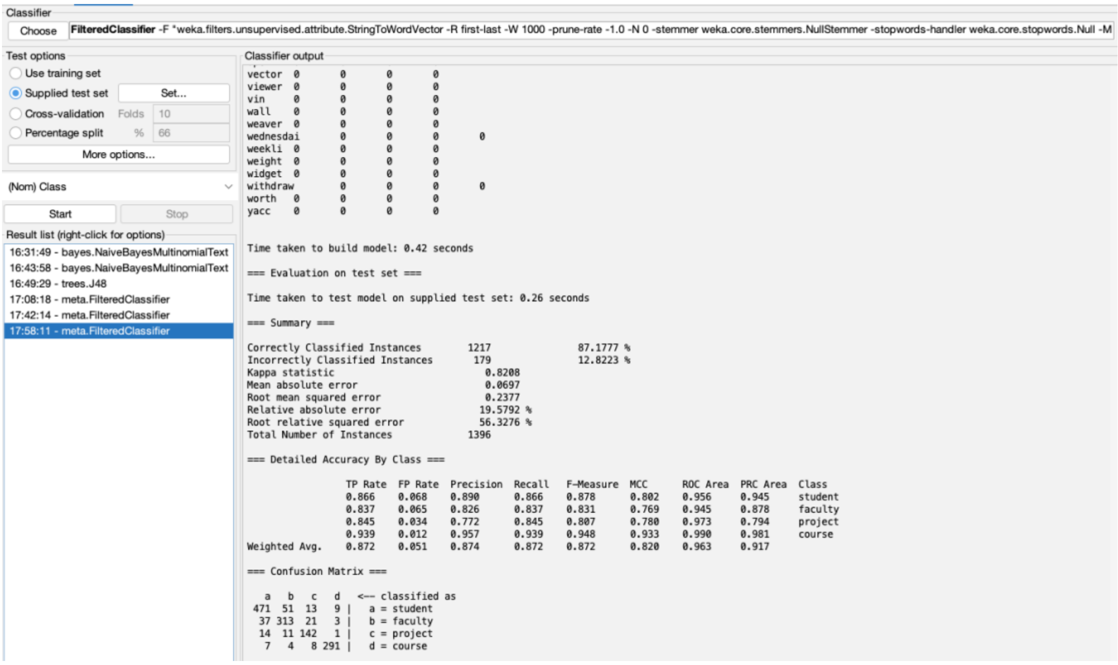
- 16:31:49 - bayes.NaiveBayesMultinomialText
- 16:43:58 - bayes.NaiveBayesMultinomialText
- 16:49:29 - trees.J48
- 17:08:18 - meta.FilteredClassifier
- 17:42:14 - meta.FilteredClassifier
- 17:58:11 - meta.FilteredClassifier
- 18:03:23 - meta.FilteredClassifier
- 18:03:53 - meta.FilteredClassifier
- 18:04:33 - meta.FilteredClassifier
- 18:05:09 - meta.FilteredClassifier
- 18:06:17 - meta.FilteredClassifier
- 18:06:27 - meta.FilteredClassifier
- 18:10:39 - meta.FilteredClassifier
- 18:11:17 - meta.FilteredClassifier
- 18:12:03 - meta.FilteredClassifier
- 18:12:21 - meta.FilteredClassifier
- 18:31:22 - meta.FilteredClassifier
- 18:31:56 - meta.FilteredClassifier
- 18:36:37 - meta.FilteredClassifier

Classifier output

```
(research <= 0) and (materi >= 1) and (boston <= 0) => Class=course (26.0/8.0)
(scienc <= 0) and (cse >= 1) and (research <= 0) => Class=course (5.0/0.0)
(professor >= 1) and (student <= 0) => Class=faculty (442.0/20.0)
(professor >= 1) and (associ >= 1) => Class=faculty (77.0/8.0)
(fax >= 1) and (work <= 0) and (advisor <= 0) and (home <= 0) => Class=faculty (83.0/20.0)
(professor >= 1) and (research >= 1) and (paper >= 1) => Class=faculty (19.0/6.0)
(faculti >= 1) and (member >= 1) => Class=faculty (21.0/3.0)
(teach >= 1) and (larg >= 1) => Class=faculty (8.0/1.0)
(lectur >= 1) and (system <= 0) => Class=faculty (15.0/4.0)
 => Class=student (1208.0/214.0)

Number of Rules : 27


Time taken to build model: 90.6 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.13 seconds

=== Summary ===

Correctly Classified Instances        1124               80.5158 %
Incorrectly Classified Instances       272               19.4842 %
Kappa statistic                          0.7243
Mean absolute error                      0.1429
Root mean squared error                  0.2833
Relative absolute error                 40.1312 %
Root relative squared error             67.1267 %
Total Number of Instances             1396

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.873    0.153    0.785      0.873    0.827      0.709    0.885     0.765     student
                 0.762    0.048    0.853      0.762    0.805      0.741    0.889     0.799     faculty
                 0.560    0.063    0.550      0.560    0.555      0.493    0.816     0.423     project
                 0.871    0.015    0.944      0.871    0.906      0.882    0.943     0.886     course
Weighted Avg.    0.805    0.083    0.810      0.805    0.806      0.730    0.891     0.760

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 475  27  36   6 |   a = student
  53 285  29   7 |   b = faculty
  55  16  94   3 |   c = project
  22   6  12 270 |   d = course
```

4. Multinomial Naïve Bayes

This classifier solves most of problems that Naïve Bayes has.
Accuracy: 87.1777% ROC Area is 0.963. And it is way faster than Naïve Bay.



Classifier

Choose **FilteredClassifier** -F "weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M

Test options

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Class

Result list (right-click for options)

- 16:31:49 - bayes.NaiveBayesMultinomialText
- 16:43:58 - bayes.NaiveBayesMultinomialText
- 16:49:29 - trees.J48
- 17:08:18 - meta.FilteredClassifier
- 17:42:14 - meta.FilteredClassifier
- 17:58:11 - meta.FilteredClassifier

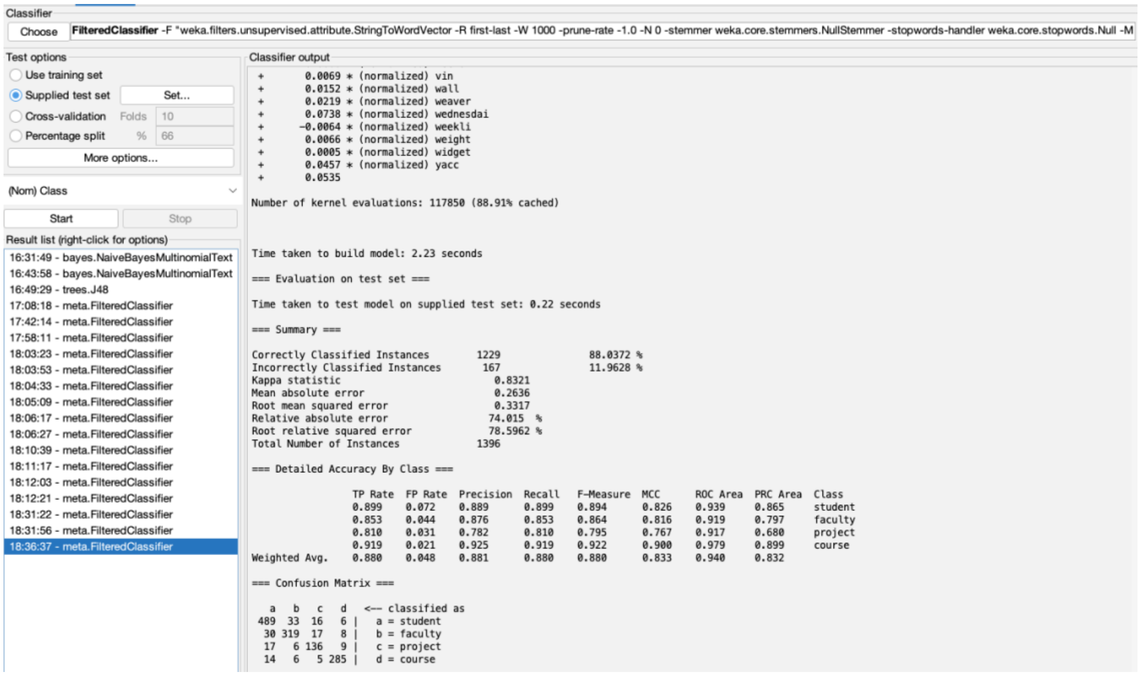
Classifier output

```
vector  0       0       0       0
viewer  0       0       0       0
vin     0       0       0       0
wall    0       0       0       0
weaver  0       0       0       0
wednesdai       0       0       0       0
weekli  0       0       0       0
weight  0       0       0       0
widget  0       0       0       0
withdraw        0       0       0       0
worth   0       0       0       0
yacc    0       0       0       0


Time taken to build model: 0.42 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.26 seconds

=== Summary ===

Correctly Classified Instances        1217               87.1777 %
Incorrectly Classified Instances       179               12.8223 %
Kappa statistic                          0.8208
Mean absolute error                      0.0697
Root mean squared error                  0.2377
Relative absolute error                 19.5792 %
Root relative squared error             56.3276 %
Total Number of Instances             1396

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.866    0.068    0.890      0.866    0.878      0.802    0.956     0.945     student
                 0.837    0.065    0.826      0.837    0.831      0.769    0.945     0.878     faculty
                 0.845    0.034    0.772      0.845    0.807      0.780    0.973     0.794     project
                 0.939    0.012    0.957      0.939    0.948      0.933    0.990     0.981     course
Weighted Avg.    0.872    0.051    0.874      0.872    0.872      0.820    0.963     0.917

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 471  51  13   9 |   a = student
  37 313  21   3 |   b = faculty
  14  11 142   1 |   c = project
   7   4   8 291 |   d = course
```

5. Sequential Minimal Optimization – SMO (training a support vector classifier)

Accuracy: 88.0372%. ROC Area is 0.94.
Deal with large feature space (high dimensional input spaces)
Assume most features are irrelevant
It can find good parameter settings automatically

Nha Tran Z23537257



Classifier

Choose **FilteredClassifier** -F "weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M

Test options

○ Use training set
● Supplied test set — Set...
○ Cross-validation — Folds: 10
○ Percentage split — %: 66

More options...

(Nom) Class

Start | Stop

Result list (right-click for options)

- 16:31:49 - bayes.NaiveBayesMultinomialText
- 16:43:58 - bayes.NaiveBayesMultinomialText
- 16:49:29 - trees.J48
- 17:08:18 - meta.FilteredClassifier
- 17:42:14 - meta.FilteredClassifier
- 17:58:11 - meta.FilteredClassifier
- 18:03:23 - meta.FilteredClassifier
- 18:03:53 - meta.FilteredClassifier
- 18:04:33 - meta.FilteredClassifier
- 18:05:09 - meta.FilteredClassifier
- 18:06:17 - meta.FilteredClassifier
- 18:06:27 - meta.FilteredClassifier
- 18:10:39 - meta.FilteredClassifier
- 18:11:17 - meta.FilteredClassifier
- 18:12:03 - meta.FilteredClassifier
- 18:12:21 - meta.FilteredClassifier
- 18:31:22 - meta.FilteredClassifier
- 18:31:56 - meta.FilteredClassifier
- 18:36:37 - meta.FilteredClassifier

Classifier output

```
 +       0.0069 * (normalized) vin
 +       0.0152 * (normalized) wall
 +       0.0219 * (normalized) weaver
 +       0.0738 * (normalized) wednesdai
 +      -0.0064 * (normalized) weekli
 +       0.0066 * (normalized) weight
 +       0.0005 * (normalized) widget
 +       0.0457 * (normalized) yacc
 +       0.0535

Number of kernel evaluations: 117850 (88.91% cached)



Time taken to build model: 2.23 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.22 seconds

=== Summary ===

Correctly Classified Instances        1229               88.0372 %
Incorrectly Classified Instances       167               11.9628 %
Kappa statistic                          0.8321
Mean absolute error                      0.2636
Root mean squared error                  0.3317
Relative absolute error                 74.015  %
Root relative squared error             78.5962 %
Total Number of Instances             1396

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.899    0.072    0.889      0.899    0.894      0.826    0.939     0.865     student
                 0.853    0.044    0.876      0.853    0.864      0.816    0.919     0.797     faculty
                 0.810    0.031    0.782      0.810    0.795      0.767    0.917     0.680     project
                 0.919    0.021    0.925      0.919    0.922      0.900    0.979     0.899     course
Weighted Avg.    0.880    0.048    0.881      0.880    0.880      0.833    0.940     0.832

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 489  33  16   6 |   a = student
  30 319  17   8 |   b = faculty
  17   6 136   9 |   c = project
  14   6   5 285 |   d = course
```