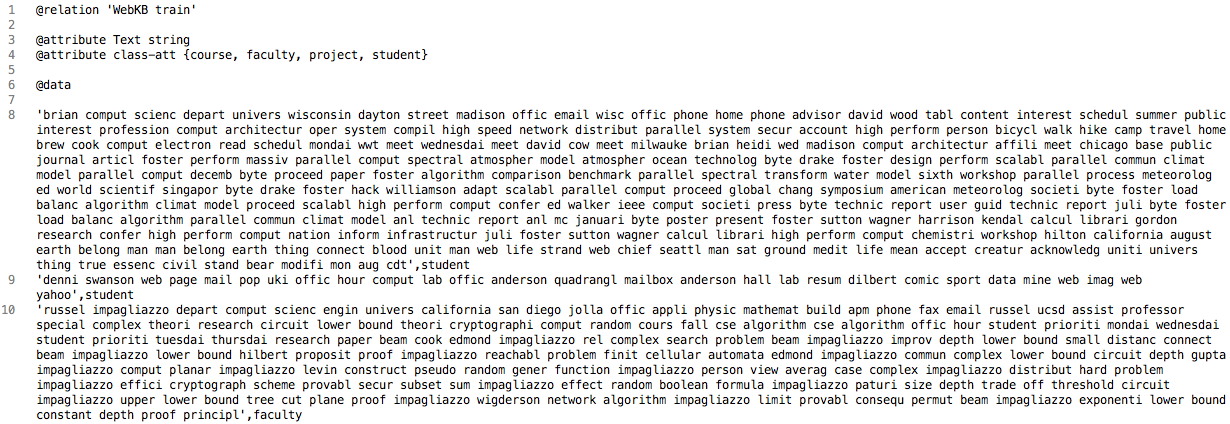
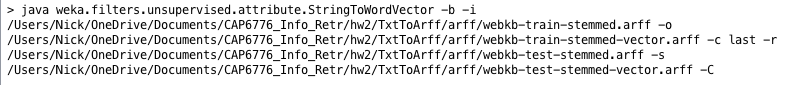
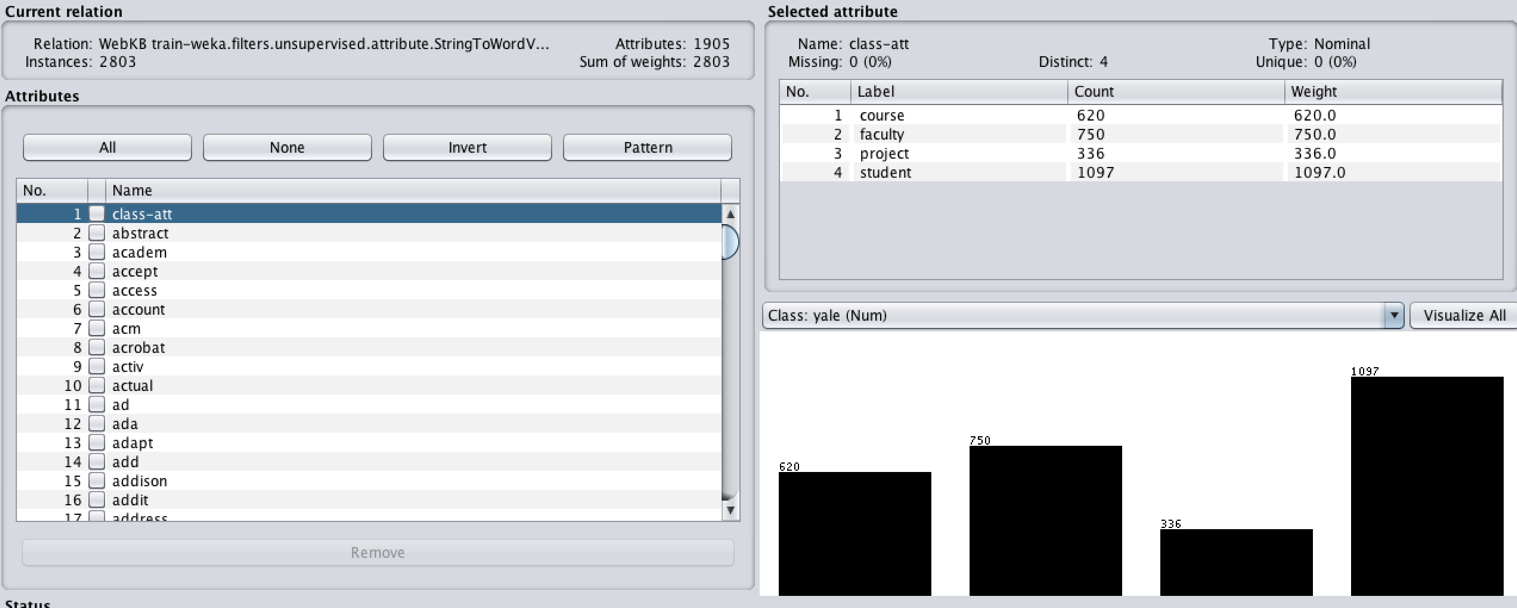
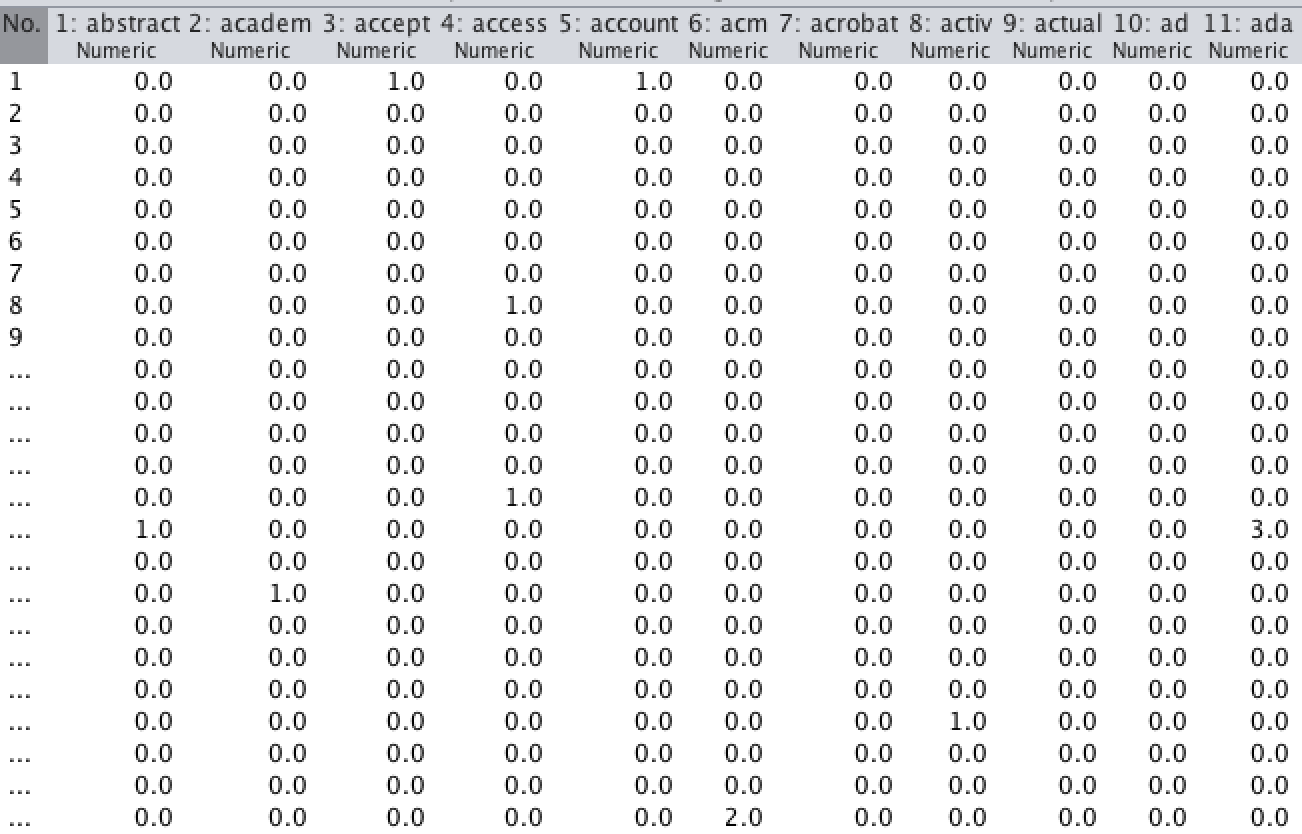
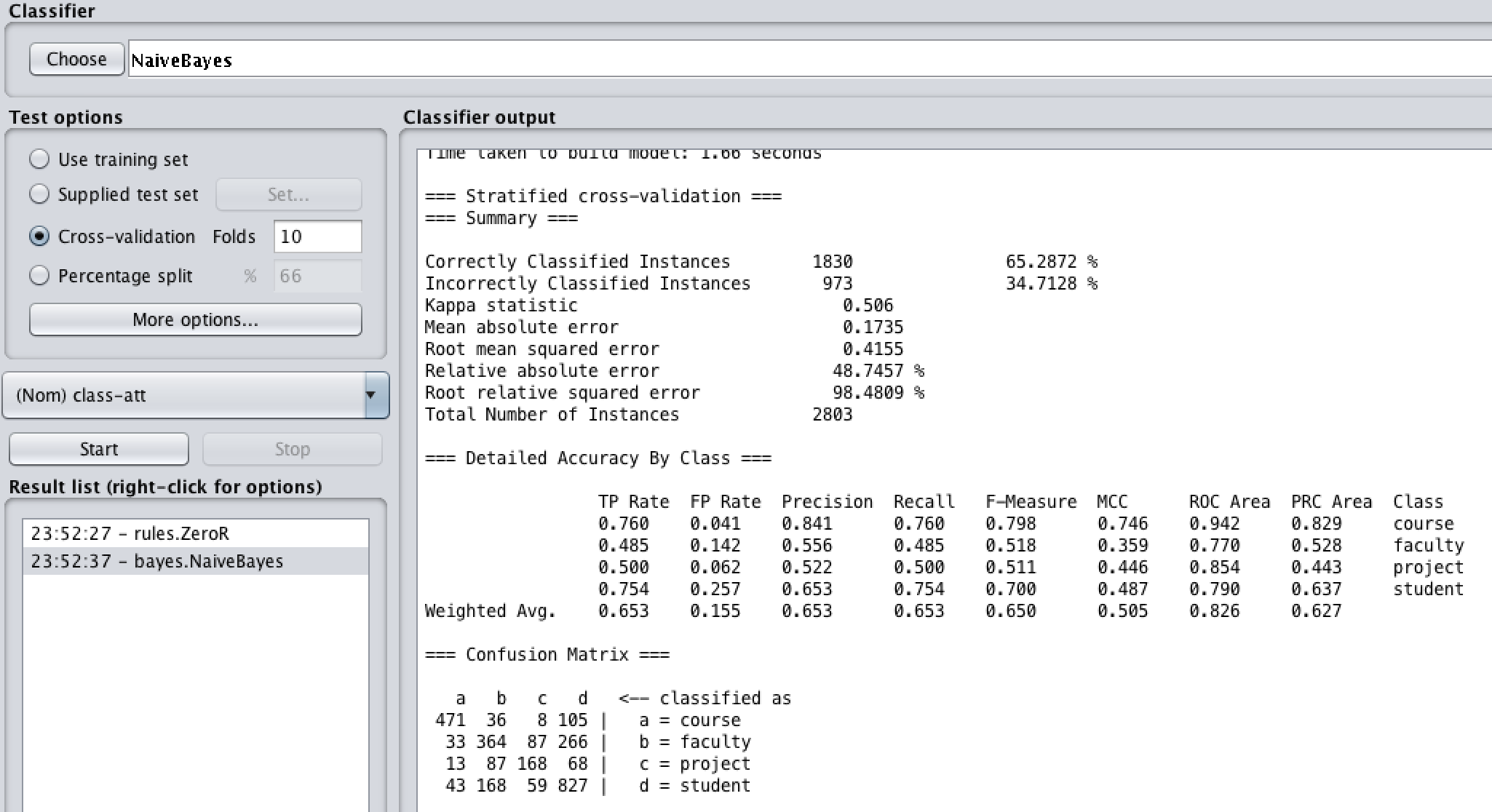
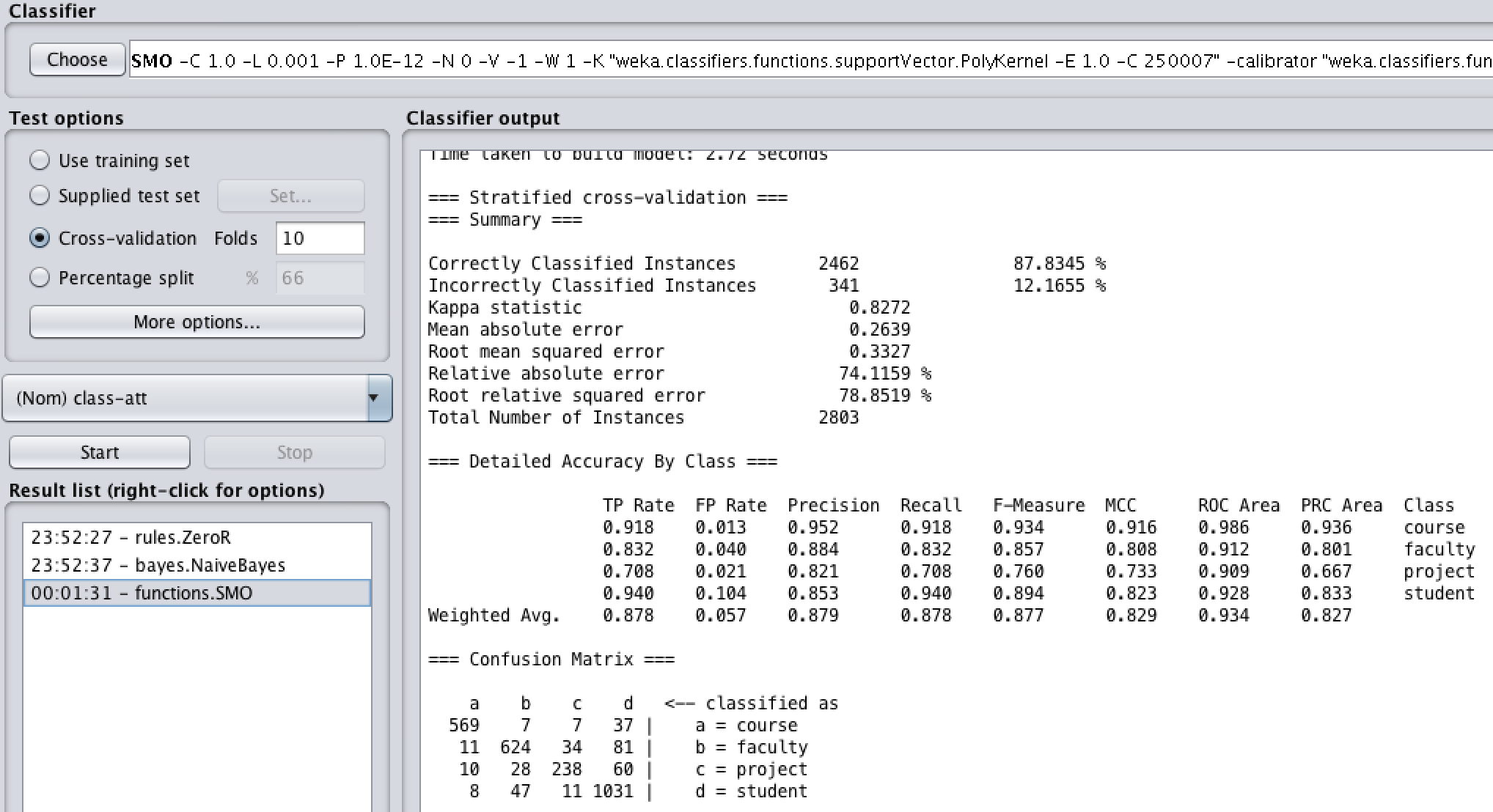
1. Converting .txt files to .arff files
   1. The provided webkb-train-stemmed.txt and webkb-test-stemmed.txt documents cannot be read by Weka.
   2. The attached Python script, TxtToArff.py, converts these files into webkb-train-stemmed.arff and webkb-test-stemmed.arff.
   3. The given files are formatted as <class>\t<document>\n, so they are converted to <.arff header>\n’<document>,<class> in the output files.
   4. Here is the header and a few documents in the training set:
2. Filtering the .arff files
   1. Using the .arff files as-is, without filtering will result in incompatible classifiers.
   2. Online tutorials recommend using the command line to apply filters to the .arff files, as shown:

This filter applies the word count parameter with the ‘-C’ argument.

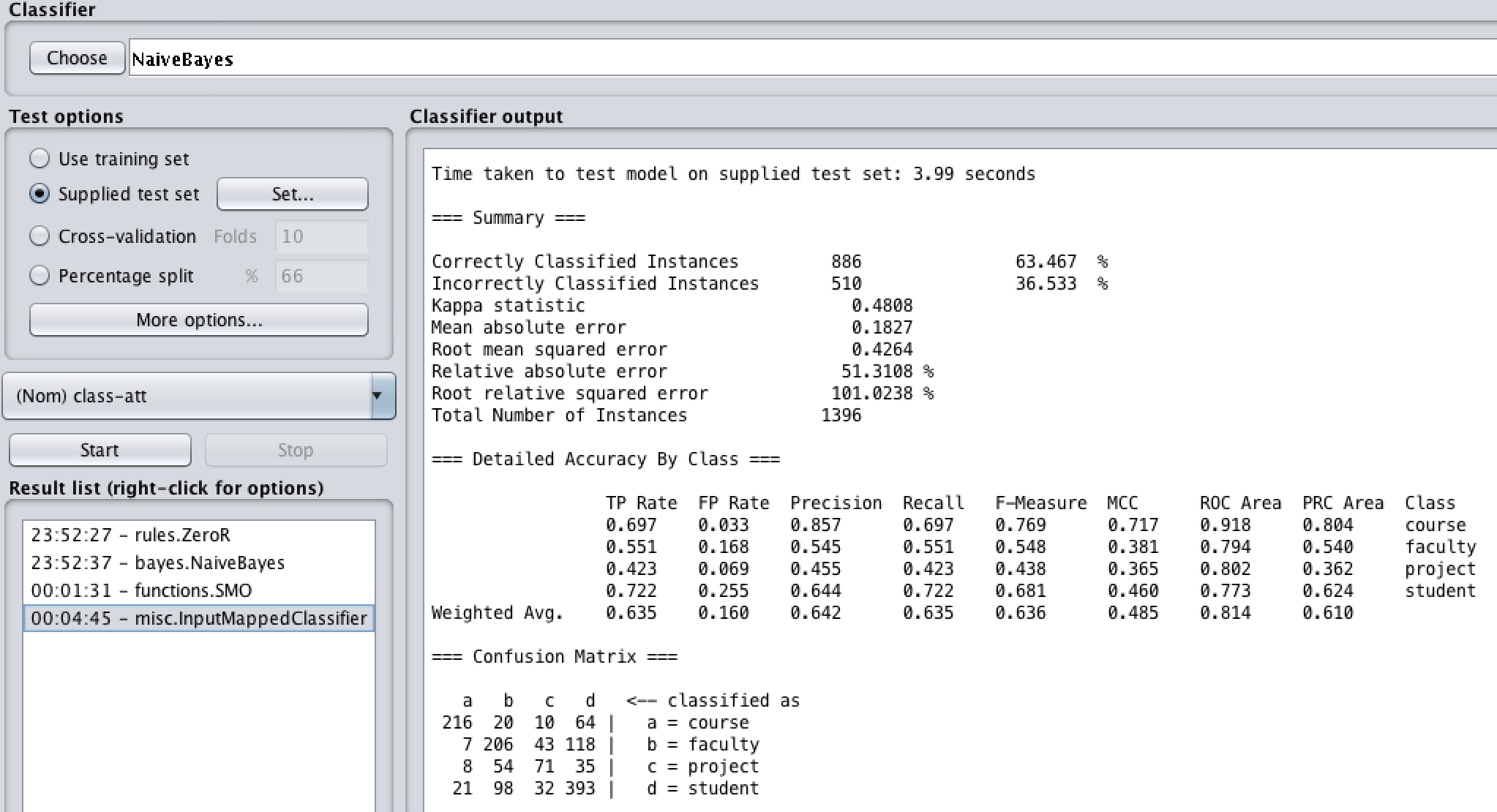
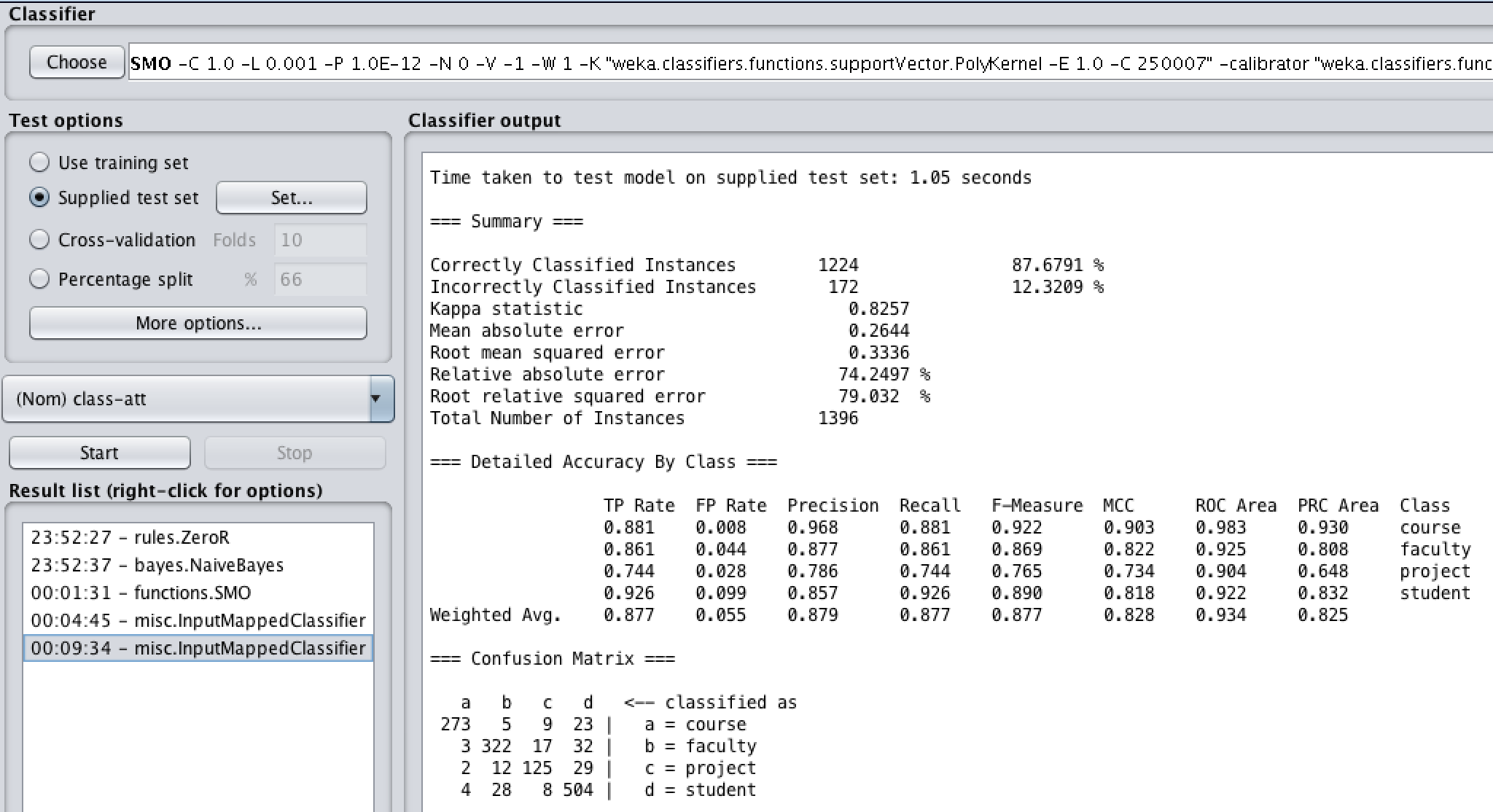
* 1. Loading the training set into Weka:
  2. Setting the class attribute as class gives a document-word matrix, which is partially shown here:



1. Building the classifiers
   1. Using the training set with Naïve Bayes and 10-fold cross validation gives the following results:
   2. The same set with SVM:



* 1. This Naïve Bayes classifier was about 65% accurate, while the Support Vector Machine classifier was about 88% accurate.
  2. For this training set, the SVM classifier has much better results.

1. Testing the classifiers
   1. The training set is run against the test set. This is with the Naïve Bayes classifier on default parameters:
   2. The training set is run against the test set with SVM and default parameters:
   3. Again, the Naïve Bayes classifier was much less accurate than the Support Vector Machine, at 64% accuracy to 88% accuracy.
   4. With the provided training and test data sets, Weka shows the SVM classification is better than Naïve Bayes at determining which documents refer to courses, students, projects, or faculty.