

Maciej Medyk – CAP6776 – Information Retrieval – Homework 02

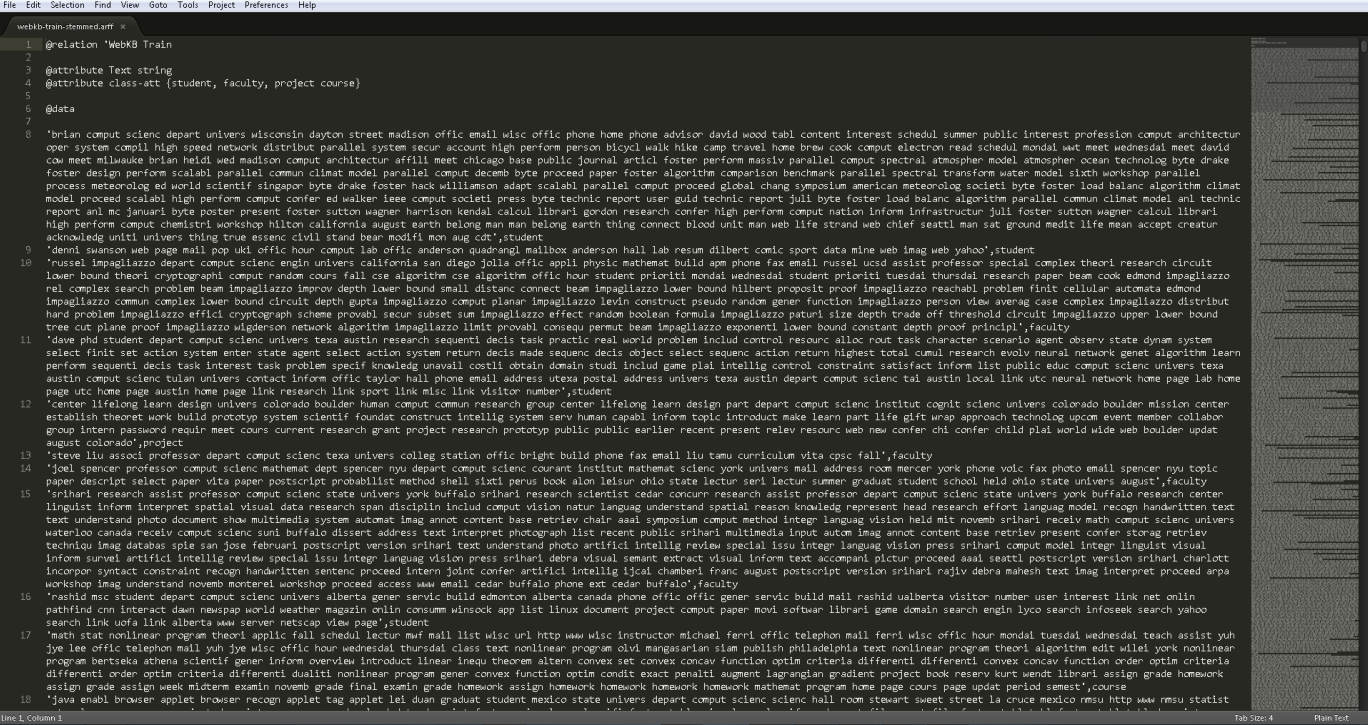
Report – [20.00pt] – Using Weka on “webkb-train-stemmed.txt” and “webkb-test-stemmed.txt”

The files received in the homework were in word format (Image 01) and could not be loaded directly to Weka. I wrote the program using C# (See Addendum) that would convert the text into string based ARFF files (Image 02). Both webkb-test-stemmed.txt and webkb-train-stemmed.txt files have been converted this way to new files webkb-test-stemmed.arff and webkb-train-stemmed.arff

Image01 (webkb-train-stemmed.txt)



Image 02 (webkb-train-stemmed.arff)



Once ARFF files were obtained it was possible to load them into WEKA where I was able to convert them further using StringToWordVector in order to create ARRF files that are vector based (Image 03 and Image 04). Initially I did boolean word count and later I did integer word count.

Image 03 (webkb-train-stemmed-vector.arff)

```

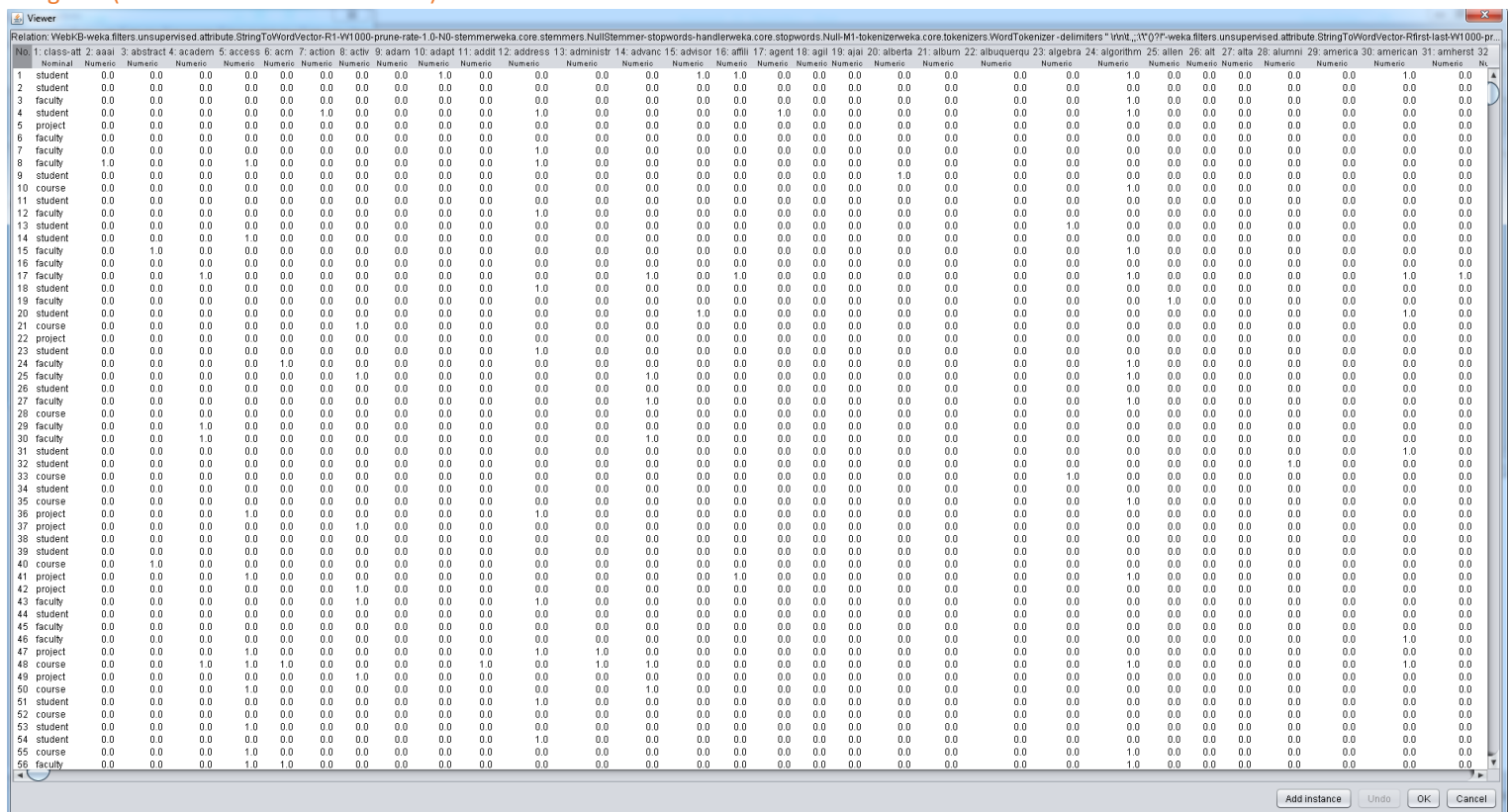
Edit Selection Find View Goto Tools File Project Preferences Help
webkb-train-stemmer-vector.affix
1 @relation 'WebKB-weka.filters.unsupervised.attribute.StringTfWordVector-R1-W1000-prune-nate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \\r\\n\\t,;|'\"'()?!\\\"'-weka.filters.unsupervised.attribute.StringTfWordVector-Rfirst-last-W1000-prune-nate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \\r\\n\\t,;|'\"'()?!\\\"'"
2
3 @attribute class attr {student,faculty,project,course}
4 @attribute aaai numeric
5 @attribute abstract numeric
6 @attribute academ numeric
7 @attribute access numeric
8 @attribute acm numeric
9 @attribute action numeric
10 @attribute activ numeric
11 @attribute adam numeric
12 @attribute adapt numeric
13 @attribute addit numeric
14 @attribute address numeric
15 @attribute administr numeric
16 @attribute advanc numeric
17 @attribute advisor numeric
18 @attribute affill numeric
19 @attribute agent numeric
20 @attribute agil numeric
21 @attribute ajal numeric
22 @attribute alberta numeric
23 @attribute album numeric
24 @attribute albuquerque numeric
25 @attribute algebra numeric
26 @attribute algorithm numeric
27 @attribute allen numeric
28 @attribute alt numeric
29 @attribute alta numeric
30 @attribute alumni numeric
31 @attribute america numeric
32 @attribute american numeric
33 @attribute amherst numeric
34 @attribute amit numeric
35 @attribute analysis numeric
36 @attribute anderson numeric
37 @attribute andrew numeric
38 @attribute anim numeric
39 @attribute annual numeric
40 @attribute anton numeric
41 @attribute appl numeric
42 @attribute applet numeric
43 @attribute appll numeric
44 @attribute applic numeric
45 @attribute approach numeric
46 @attribute april numeric
47 @attribute apt numeric
48 @attribute architectur numeric
49 @attribute archiv numeric
50 @attribute area numeric

```

Image 04 (webkb-test-stemmed-vector.arff)

```
File Edit View Find View Goto Tools Window Help
webkit-stemmed-vector.affix
1 |relation 'Web8-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.
   tokenizers.WordTokenizer -delimiters "\r\n\t.,;:\\'\"(){}|`~"
2
3 @attribute class::att {student,faculty,project,course}
4 @attribute aaal numeric
5 @attribute abil numeric
6 @attribute abstract numeric
7 @attribute academ numeric
8 @attribute accept numeric
9 @attribute access numeric
10 @attribute achiev numeric
11 @attribute acl numeric
12 @attribute acm numeric
13 @attribute acouisit numeric
14 @attribute action numeric
15 @attribute activ numeric
16 @attribute ad numeric
17 @attribute adam numeric
18 @attribute adapt numeric
19 @attribute addit numeric
20 @attribute address numeric
21 @attribute adminisr numeric
22 @attribute advanc numeric
23 @attribute advisor numeric
24 @attribute affill numeric
25 @attribute agent numeric
26 @attribute aid numeric
27 @attribute alan numeric
28 @attribute alberta numeric
29 @attribute album numeric
30 @attribute algebra numeric
31 @attribute algo-rithm numeric
32 @attribute alloc numeric
33 @attribute altavista numeric
34 @attribute alumni numeric
35 @attribute america numeric
36 @attribute american numeric
37 @attribute amherst numeric
38 @attribute analog numeric
39 @attribute analysi numeric
40 @attribute andrew numeric
41 @attribute anim numeric
42 @attribute annual numeric
43 @attribute appear numeric
44 @attribute appl numeric
45 @attribute applet numeric
46 @attribute appll numeric
47 @attribute apple numeric
48 @attribute approach numeric
49 @attribute approxim numeric
50 @attribute april numeric
51 @attribute apt numeric
```

Image 05 (webkb-train-stemmed-vector.arff loaded into Weka)



Once I had both files converted I was able to run Naïve Bayes classification using 10-fold cross-validation (Image 07). I was able to obtain results in Image 08 and accuracy was calculated to 76.49 % ((885 + 471 + 247 + 541) / 2803).

Image 07 (running NaïveBayes classifier using 10-fold cross validation having Boolean word count)

Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Classifier

ChooseNaiveBayes

Test options

Use training set

Supplied test set

Cross-validation

Folds10

Percentage split

%66

More options...

(Nom) class-all

StartStop

Result list (right-click for options)

21:48:32 - bayes.NaiveBayes

Classifier output

Run information
Schema: weka.classifiers.bayes.NaiveBayes
Relation: WebKB-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-NO-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.NullN1-tokenizerweka.core.tokenizers.WordTokenizer
Instances: 2803
Attributes: 1905
Test mode: 10-fold cross-validation
Classifier model (full training set)
Naive Bayes Classifier
AttributeClass
student(0.39)(0.27)(0.12)(0.22)
faculty0.01550.04270.00890
project0.16670.20210.16670.1667
course0.10970.7500.3360.620
precision1111
abstract0.04380.06930.1220.0645
std. dev.0.20460.2540.32730.2457
weight sum1097750336620
precision1111
academ0.06930.11330.03570.0919
std. dev.0.25390.3170.18560.2889
weight sum1097750336620
precision1111
access0.08930.0760.20830.1194
std. dev.0.28520.2650.40610.3242
weight sum1097750336620
precision1111
cm0.05290.1880.06250.0468
std. dev.0.22380.39070.24210.2112
weight sum1097750336620
precision1111

Status

OKLogx0

Image 08 (results of NaïveBayes classifier using 10-fold cross validation having Boolean word count)

Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Classifier

ChooseNaiveBayes

Test options

Use training set

Supplied test set

Cross-validation

Folds10

Percentage split

%66

More options...

(Nom) class-all

StartStop

Result list (right-click for options)

21:48:32 - bayes.NaiveBayes

Classifier output

worth
mean0.00910.00130.0030.0177
std. dev.0.16670.16670.16670.1667
weight sum1097750336620
precision1111
yacc
mean0.003600.00129
std. dev.0.16670.16670.16670.1667
weight sum1097750336620
precision1111
Time taken to build model: 2.01 seconds
Stratified cross-validation
Summary
Correctly Classified Instances214476.4895 %
Incorrectly Classified Instances65923.5105 %
Kappa statistic0.6704
Mean absolute error0.1178
Root mean squared error0.335
Relative absolute error33.1019 %
Root relative squared error79.413 %
Total Number of Instances2803
Detailed Accuracy By Class
TP RateFP RatePrecisionRecallF-MeasureMCCROC AreaPRC AreaClass
0.8070.1710.7530.8070.7790.6290.8990.880student
0.6280.0700.7670.6280.6910.5980.8910.777faculty
0.7350.0810.5530.7350.6310.5800.9340.630project
0.8730.0110.9560.8730.9120.8900.9860.966course
Weighted Avg.0.7650.0980.7770.7650.7670.6730.9200.841
Confusion Matrix
a b c d <-- classified as
885 114 83 15 | a = student
173 471 98 8 | b = faculty
63 24 247 2 | c = project
55 5 19 541 | d = course

Status

OKLogx0

I also run Naïve Bayes classification while supplying test set (Image 09). I was able to obtain results in Image 10 and accuracy was calculated to 77.87 % ((424 + 258 + 139 + 266) / 1396).

Image 09 (running NaïveBayes classifier using test set file webkb-test-stemmed-vector.arff having Boolean word count in both files)

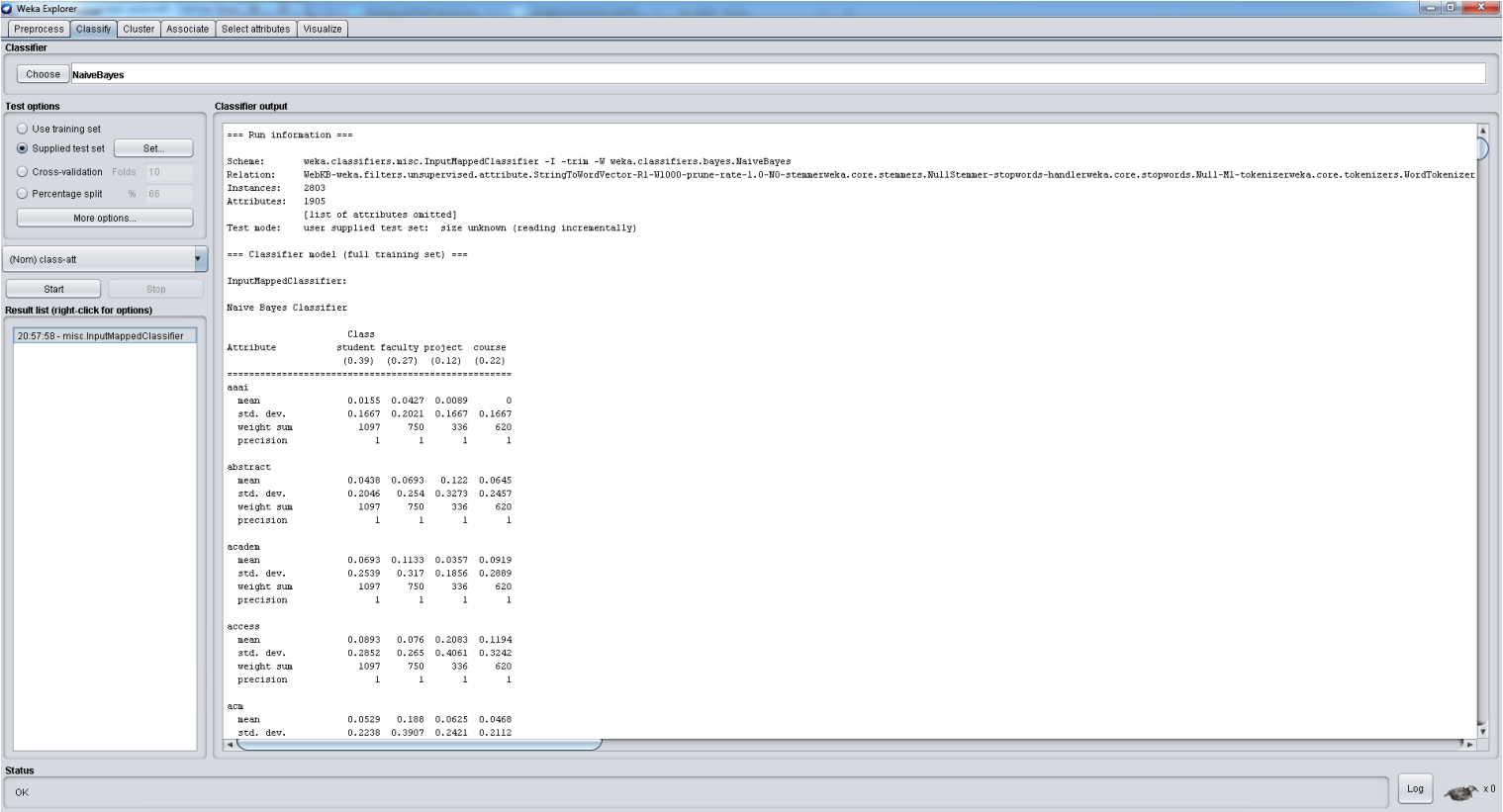
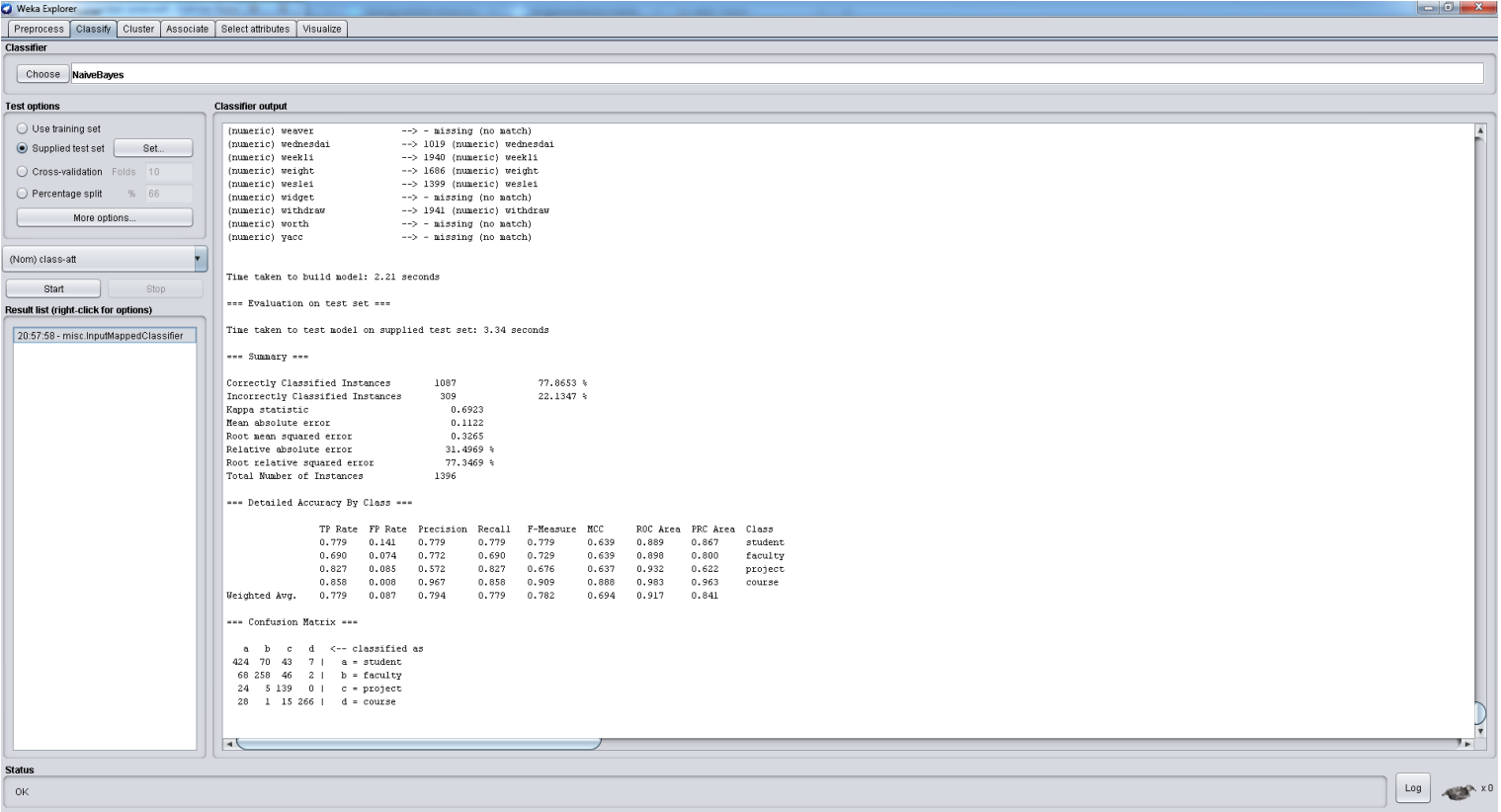


Image 10 (results of NaïveBayes classifier using test set file webkb-test-stemmed-vector.arff having Boolean word count in both files)



Afterwards, I had to install LibSVM as it was an optional package in Weka (Image 11). Then I was able to run LibSVM classification using 10-fold cross-validation (Image 12). I was able to obtain results and accuracy was calculated to 73.56 % ((1078 + 489 + 29 + 466) / 2803).

Image 11 (LibSVM package library being installed)

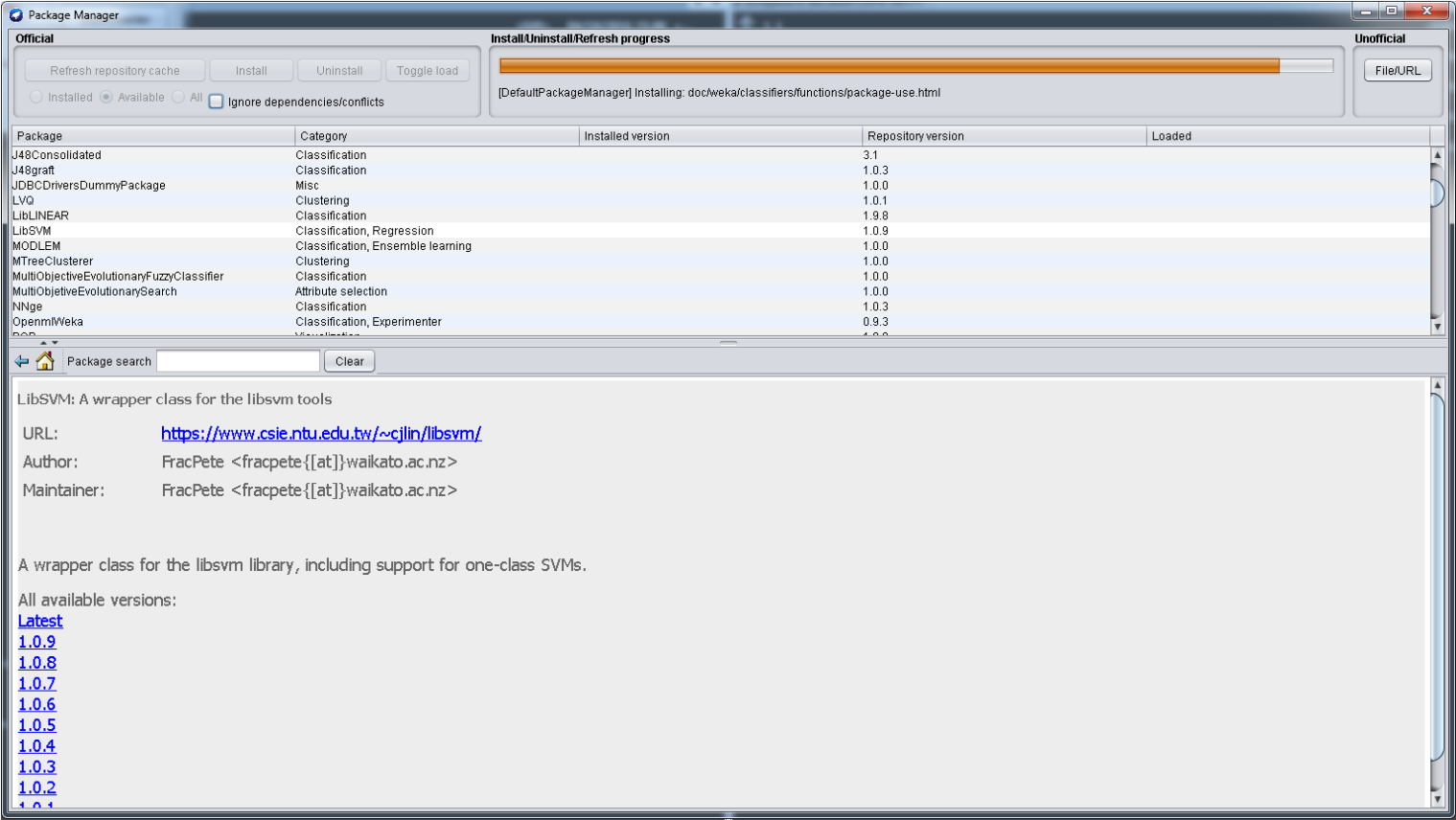
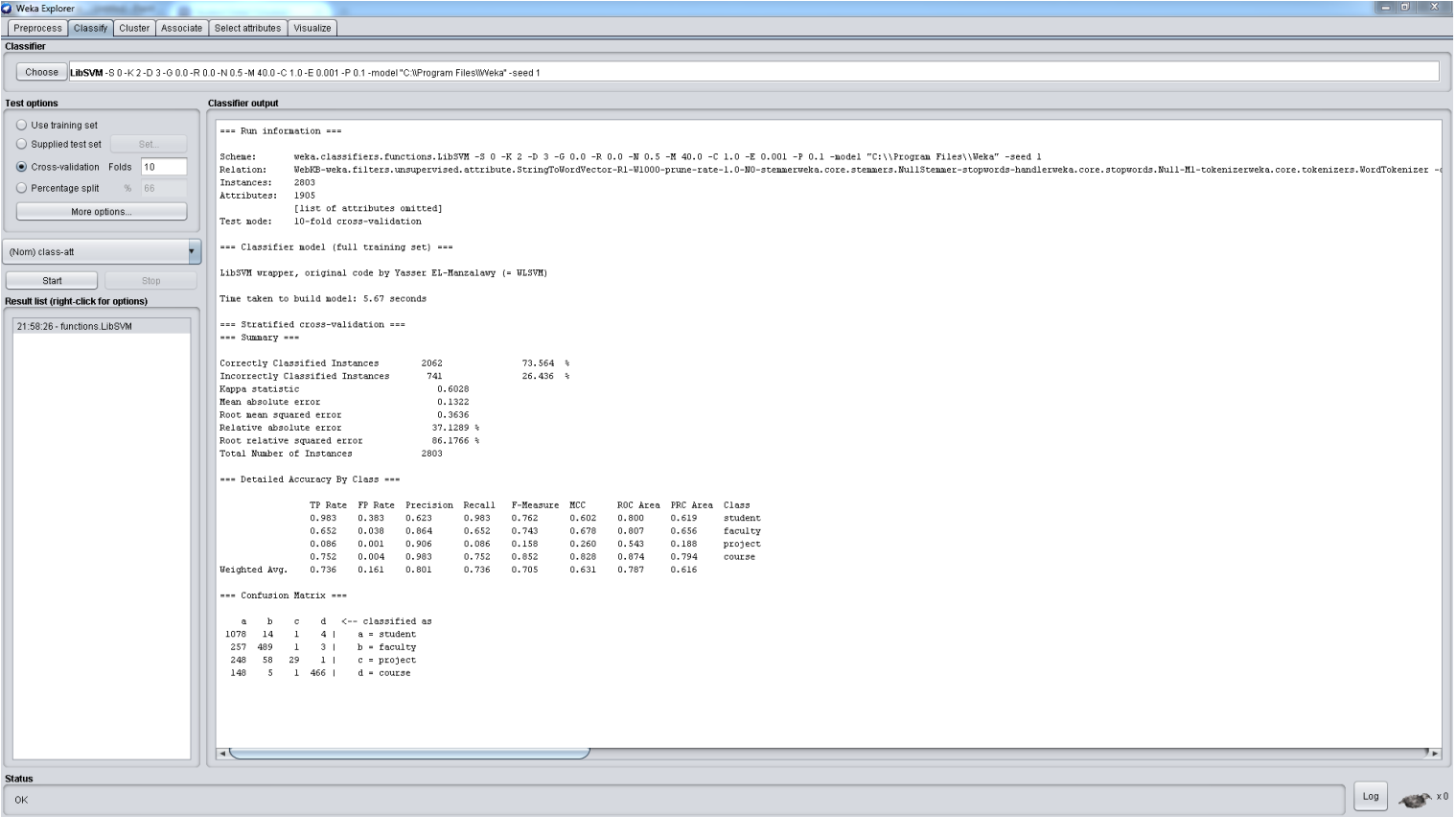


Image 12 (results Support Vector Machine classifier using 10-fold cross validation having Boolean word count)



I also run LibSVM classification while supplying test set (Image 13). I was able to obtain results in Image 14 and accuracy was calculated to 76.07 % ((526 + 279 + 26 + 231) / 1396).

Image 13 (running Support Vector Machine classifier using test set file webkb-test-stemmed-vector.arff having Boolean word count)

Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Classifier

ChooseLibSVM-S 0-K 2-D 3-G 0.0-R 0.0-N 0.5-M 40.0-C 1.0-E 0.001-P 0.1-model"C:\Program Files\Weka"-seed 1

Test options

Use training set

Supplied test setSet...

Cross-validationFolds10

Percentage split%66

More options...

(Nom) class-att

StartStop

Result list (right-click for options)

21.05.43 - misc.InputMappedClassifier

Classifier output

Run information
Scheme: weka.classifiers.misc.InputMappedClassifier -I -t x m -W weka.classifiers.functions.LibSVM -- -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -model "C:\Program Files\Weka" -seed 1
Relation: webkb-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-ML-tokenizerweka.core.tokenizers.WordTokenizer
Instances: 2803
Attributes: 1905
[list of attributes omitted]
Test mode: user supplied test set: size unknown (reading incrementally)

Classifier model (full training set) ===

InputMappedClassifier:

LibSVM wrapper, original code by Yasser El-Mansalawy (= WLSVM)
Attribute mappings:

Model attributesIncoming attributes

(nominal) class-att--> 1 (nominal) class-att
(numeric) aaai--> 2 (numeric) aaai
(numeric) abstract--> 4 (numeric) abstract
(numeric) academa--> 5 (numeric) academa
(numeric) access--> 7 (numeric) access
(numeric) acm--> 10 (numeric) acm
(numeric) action--> 12 (numeric) action
(numeric) activ--> 13 (numeric) activ
(numeric) adam--> 15 (numeric) adam
(numeric) adapt--> 16 (numeric) adapt
(numeric) addit--> 17 (numeric) addit
(numeric) address--> 18 (numeric) address
(numeric) administr--> 19 (numeric) administr
(numeric) advanc--> 20 (numeric) advanc
(numeric) advisor--> 21 (numeric) advisor
(numeric) affili--> 22 (numeric) affili
(numeric) agent--> 23 (numeric) agent
(numeric) agil--> - missing (no match)
(numeric) ajal--> - missing (no match)
(numeric) alberta--> 26 (numeric) alberta
(numeric) album--> 27 (numeric) album
(numeric) albuquerque--> - missing (no match)
(numeric) algebra--> 28 (numeric) algebra
(numeric) algorithm--> 29 (numeric) algorithm
(numeric) allen--> 1695 (numeric) allen
(numeric) alt--> - missing (no match)
(numeric) alta--> - missing (no match)
(numeric) alumni--> 32 (numeric) alumni

Status

OK

Log

x 0

Image 14 (results of Support Vector Machine classifier using test set file webkb-test-stemmed-vector.arff having Boolean word count)

Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Classifier

ChooseLibSVM-S 0-K 2-D 3-G 0.0-R 0.0-N 0.5-M 40.0-C 1.0-E 0.001-P 0.1-model"C:\Program Files\Weka"-seed 1

Test options

Use training set

Supplied test setSet...

Cross-validationFolds10

Percentage split%66

More options...

(Nom) class-att

StartStop

Result list (right-click for options)

21.05.43 - misc.InputMappedClassifier

Classifier output

(numeric) weaver--> - missing (no match)
(numeric) webnewsdai--> 1019 (numeric) webnewsdai
(numeric) weekli--> 1940 (numeric) weekli
(numeric) weight--> 1686 (numeric) weight
(numeric) weslei--> 1399 (numeric) weslei
(numeric) widget--> - missing (no match)
(numeric) withdraw--> 1941 (numeric) withdraw
(numeric) worth--> - missing (no match)
(numeric) yacc--> - missing (no match)

Time taken to build model: 6.07 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 6.53 seconds

=== Summary ===

Correctly Classified Instances106276.0748 %
Incorrectly Classified Instances33423.9255 %
Kappa statistic0.644
Mean absolute error0.1196
Root mean squared error0.3459
Relative absolute error33.5883 %
Root relative squared error81.9462 %
Total Number of Instances1396

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------|
| 0.967 | 0.335 | 0.649 | 0.967 | 0.776 | 0.625 | 0.816 | 0.640 | student | |
| 0.746 | 0.043 | 0.964 | 0.746 | 0.801 | 0.738 | 0.851 | 0.712 | faculty | |
| 0.155 | 0.003 | 0.867 | 0.155 | 0.263 | 0.340 | 0.576 | 0.236 | project | |
| 0.745 | 0.001 | 0.996 | 0.745 | 0.852 | 0.831 | 0.872 | 0.799 | course | |
| Weighted Avg. | 0.761 | 0.142 | 0.810 | 0.761 | 0.738 | 0.667 | 0.809 | 0.646 | |

=== Confusion Matrix ===

| | a | b | c | d | <-- classified as |
|-----|-----|----|-----|---|-------------------|
| 526 | 18 | 0 | 0 | 1 | a = student |
| 93 | 279 | 2 | 0 | 1 | b = faculty |
| 119 | 22 | 26 | 1 | 1 | c = project |
| 73 | 4 | 2 | 231 | 1 | d = course |

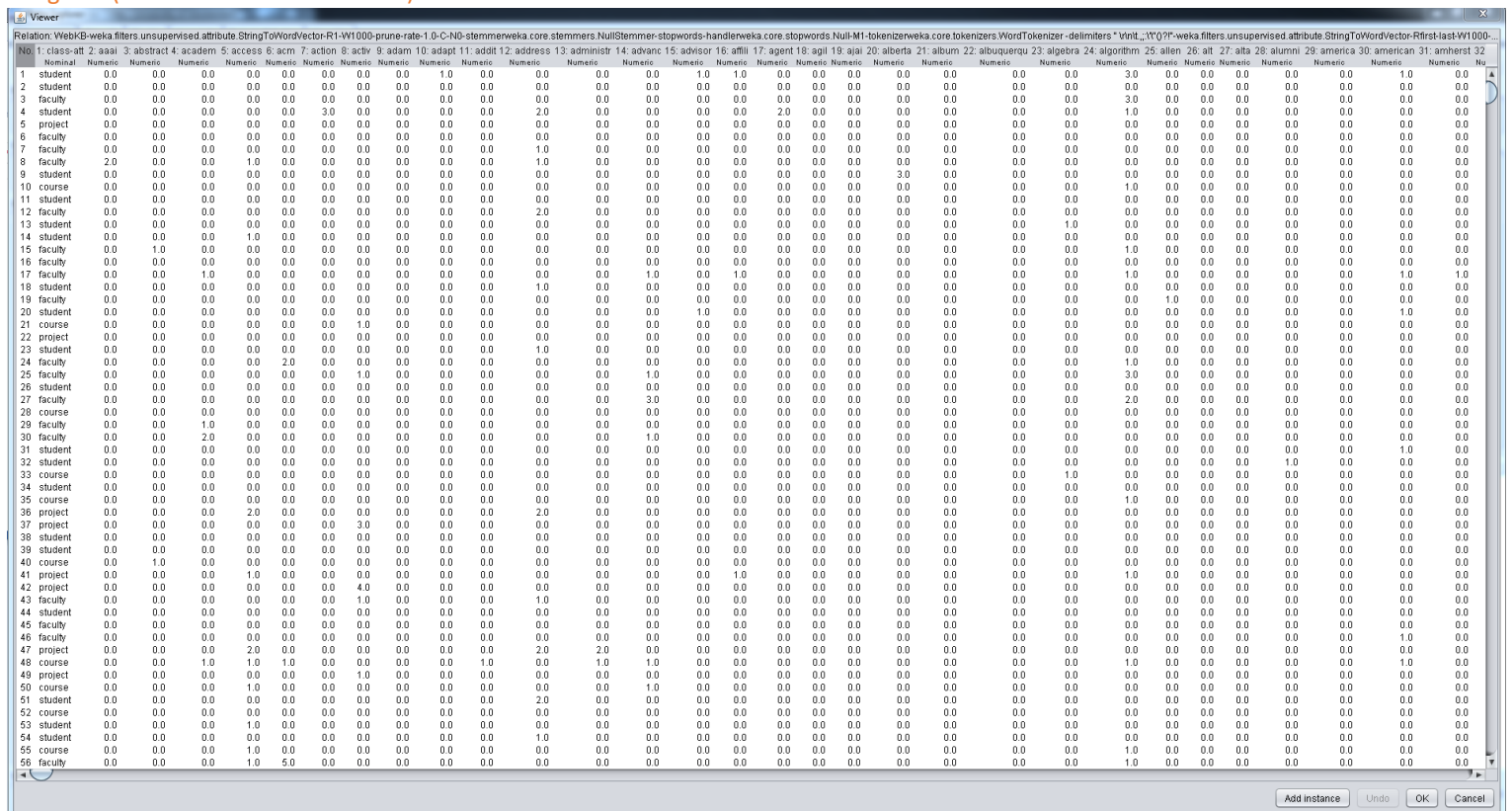
Status

OK

Log

x 0

Image 15 (webkb-train-stemmed.arff loaded into Weka having actual integer word count)



Once I had both files converted I was able to run Naïve Bayes classification using 10-fold cross-validation (Image 17). I was able to obtain results in Image 18 and accuracy was calculated to 65.39 % ((828 + 365 + 169 + 471) / 2803).

Image 17 (running NaïveBayes classifier using 10-fold cross validation having actual integer word count)

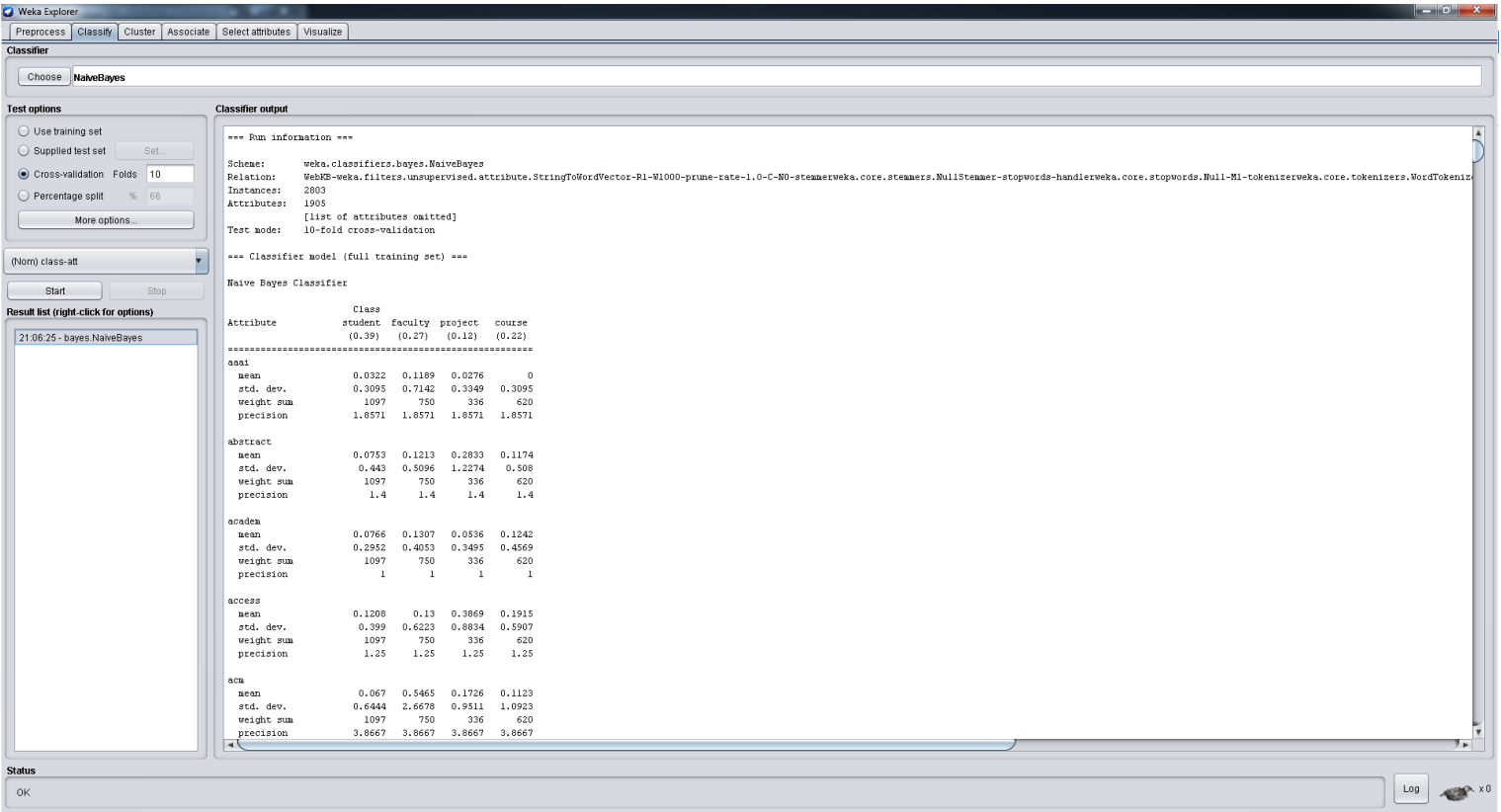
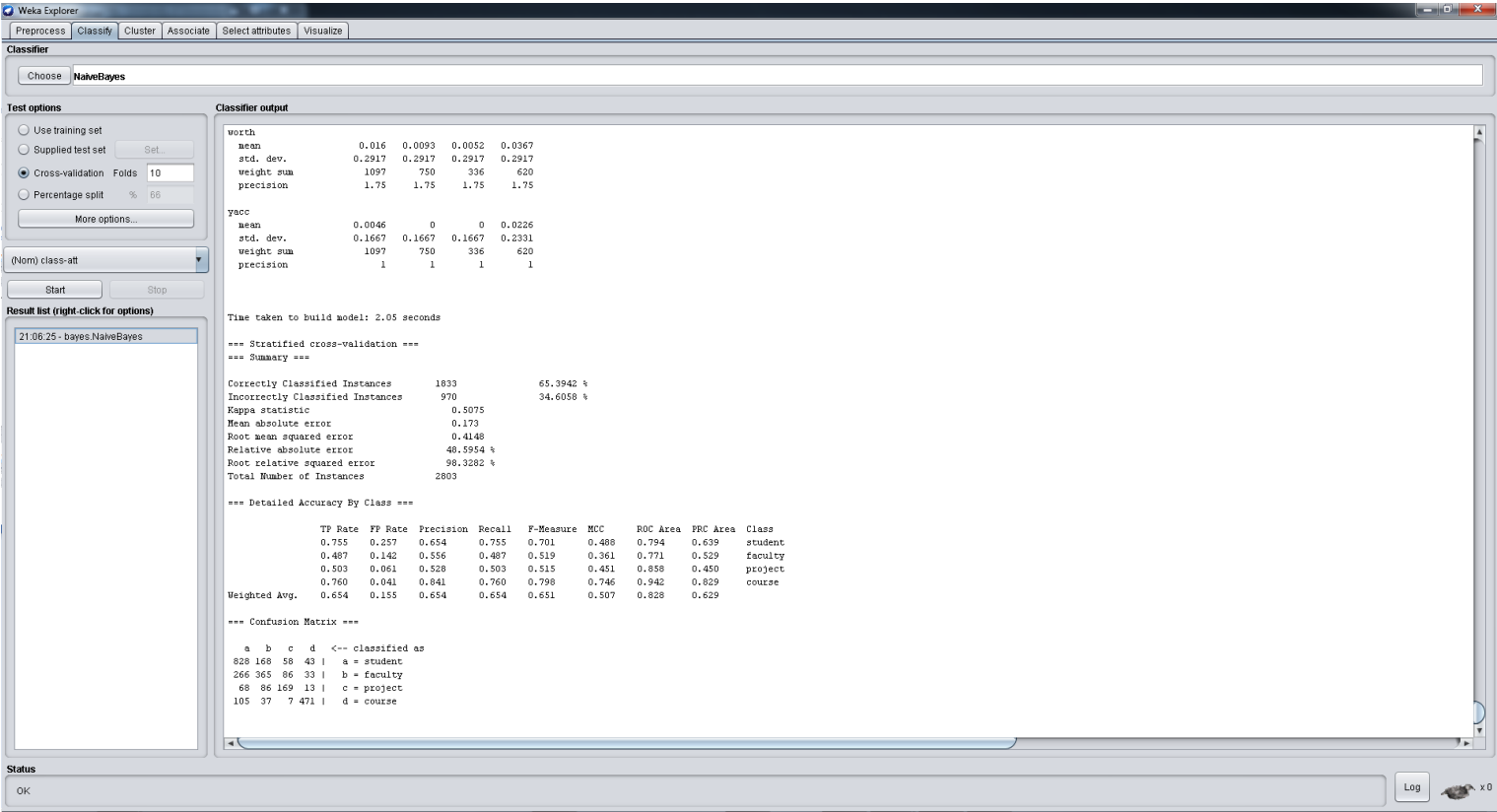


Image 18 (results of NaïveBayes classifier using 10-fold cross validation having actual integer word count)



I also run Naïve Bayes classification while supplying test set (Image 19). I was able to obtain results in Image 20 and accuracy was calculated to 63.40 % ((395 + 204 + 41 + 215) / 1396).

Image 19 (running NaïveBayes classifier using test set file webkb-test-stemmed-vector.arff having actual integer word count)

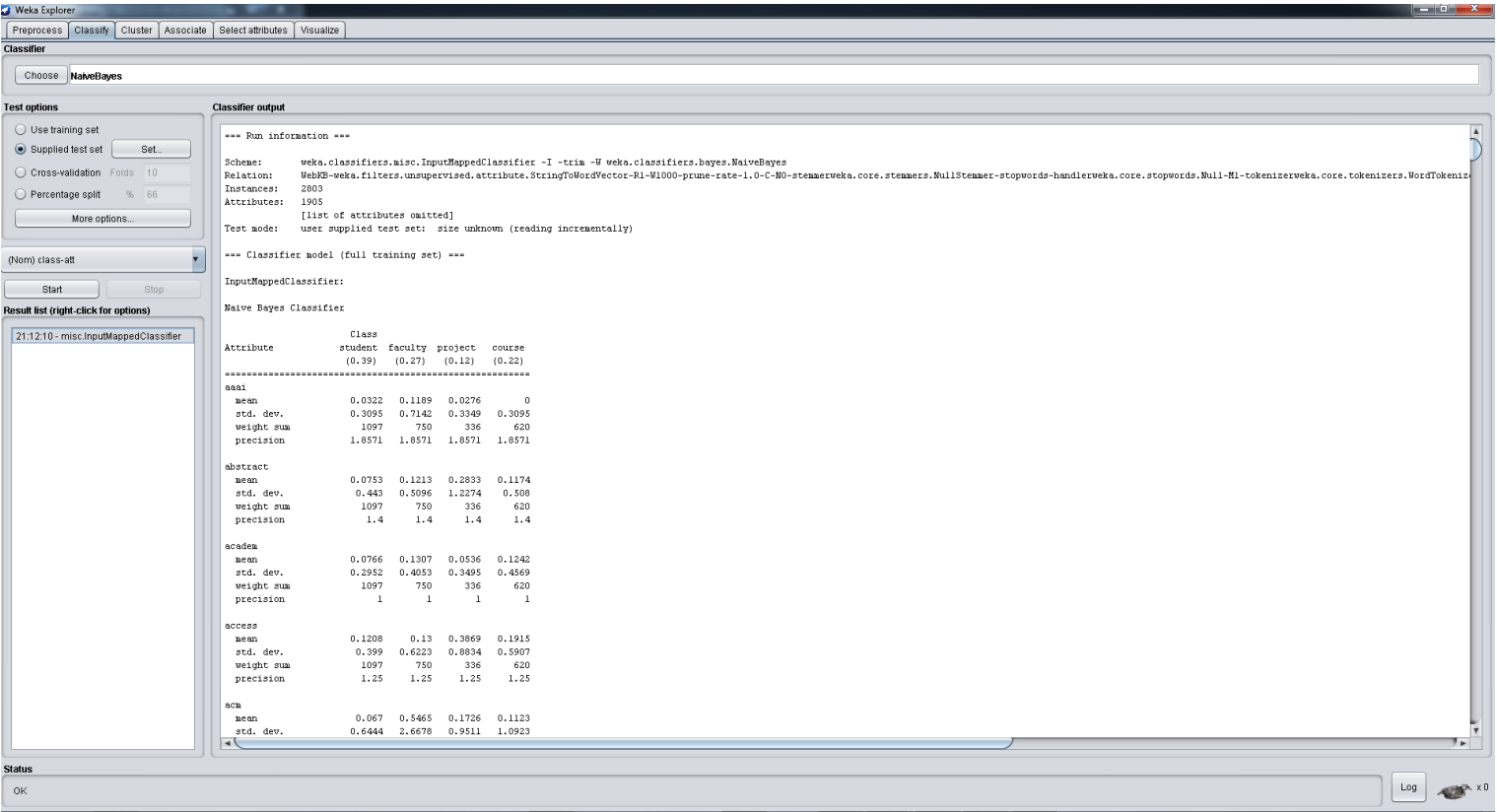
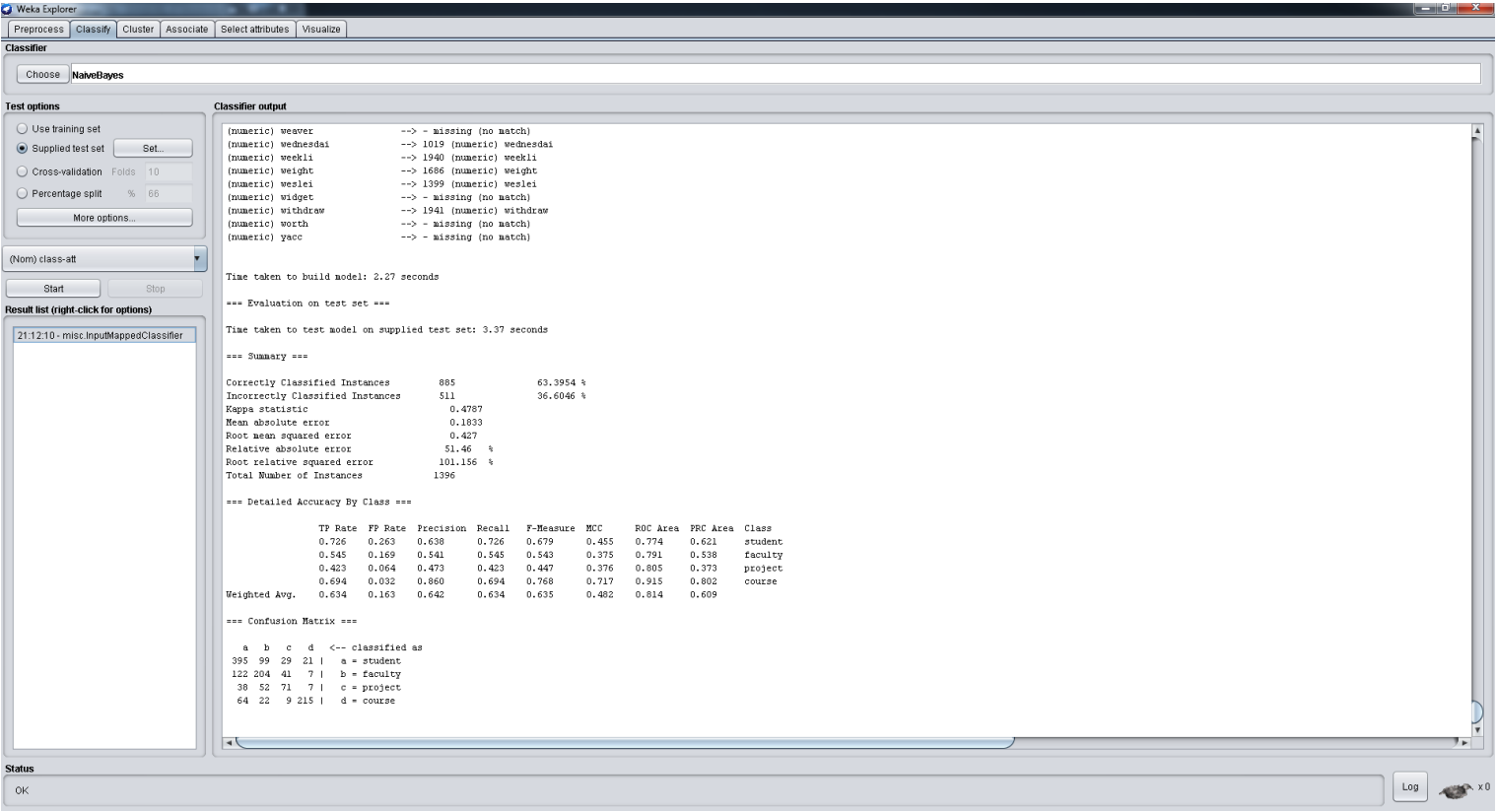
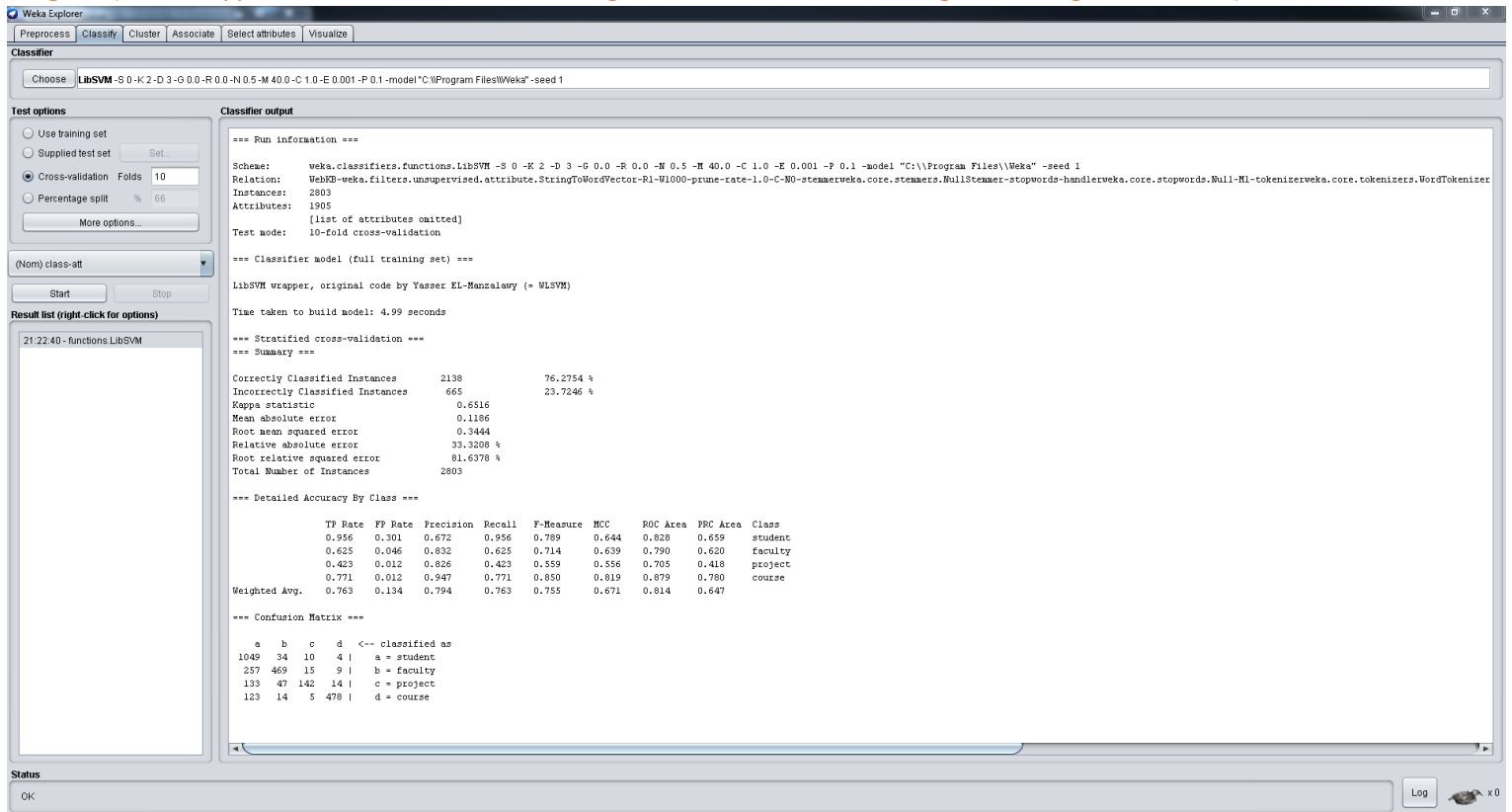


Image 20 (results of NaïveBayes classifier using test set file webkb-test-stemmed-vector.arff having actual integer word count)



Afterwards, I run LibSVM classification using 10-fold cross-validation (Image 21). I was able to obtain results and accuracy was calculated to 76.28 % ((1049 + 469 + 142 + 478) / 2803).

Image 21 (results Support Vector Machine classifier using 10-fold cross validation having actual integer word count)



I also run LibSVM classification while supplying test set (Image 22). I was able to obtain results in Image 23 and accuracy was calculated to 78.72 % ((515 + 267 + 84 + 233) / 1396).

Image 22 (running Support Vector Machine classifier using test set file webkb-test-stemmed-vector.arff having actual integer word count)

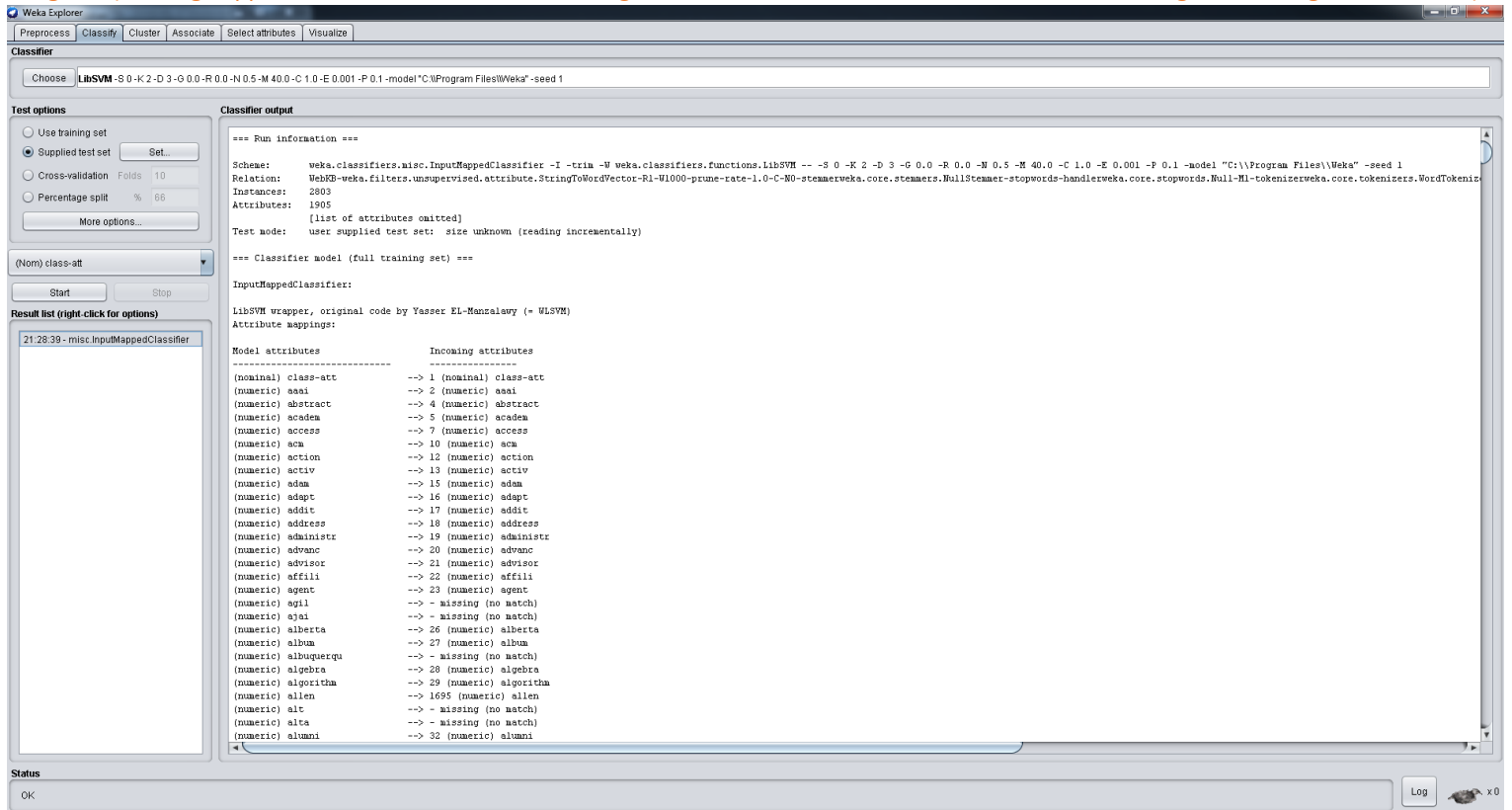
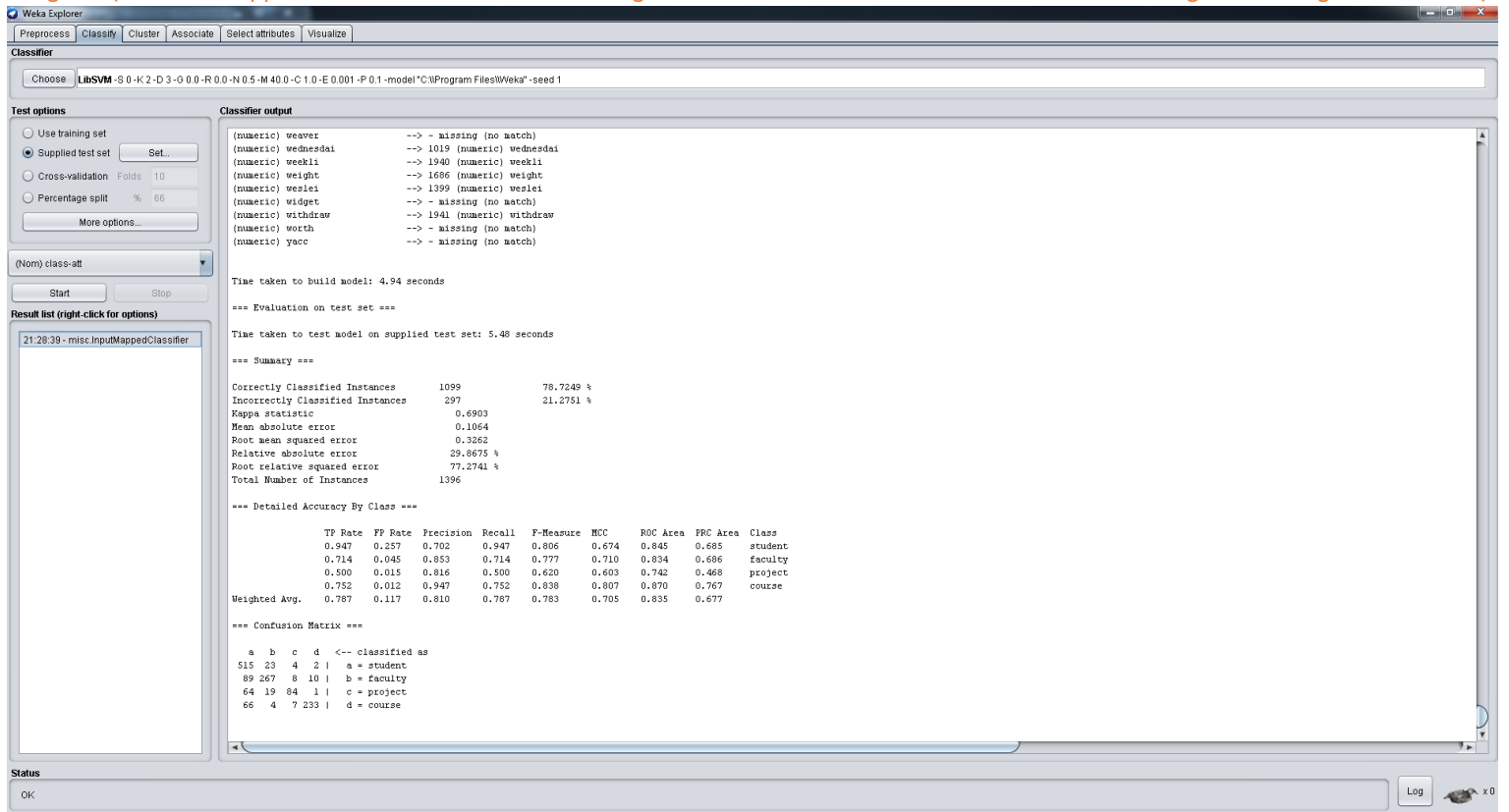


Image 23 (results of Support Vector Machine classifier using test set file webkb-test-stemmed-vector.arff having actual integer word count)



After the analysis of both Naïve Bayes and Support Vector Machine classifiers we can see that SVM while supplying test set file having actual word count was the most accurate method with accuracy of 78.28 %.

| Naïve Bayes Using Test Set Having Boolean Word Count | | | | |
|--|--|-----|-----|-----|
| | A | B | C | D |
| A | 424 | 70 | 43 | 7 |
| B | 68 | 258 | 46 | 2 |
| C | 24 | 5 | 139 | 0 |
| D | 28 | 1 | 15 | 266 |
| Accuracy | (424 + 258 + 139 + 266) / 1396 = 77.87 % | | | |

| Naïve Bayes Using 10-Fold Cross-Validation Having Boolean Word Count | | | | |
|--|--|-----|-----|-----|
| | A | B | C | D |
| A | 885 | 114 | 83 | 15 |
| B | 173 | 471 | 98 | 8 |
| C | 63 | 24 | 247 | 2 |
| D | 55 | 5 | 19 | 541 |
| Accuracy | (885 + 471 + 247 + 541) / 2803 = 76.49 % | | | |

| Naïve Bayes Using Test Set Having Actual Word Count | | | | |
|---|---|-----|----|-----|
| | A | B | C | D |
| A | 395 | 99 | 29 | 21 |
| B | 122 | 204 | 41 | 7 |
| C | 38 | 52 | 71 | 7 |
| D | 64 | 22 | 9 | 215 |
| Accuracy | (395 + 204 + 71 + 215) / 1396 = 63.40 % | | | |

| Naïve Bayes Using 10-Fold Cross-Validation Having Boolean Word Count | | | | |
|--|--|-----|-----|-----|
| | A | B | C | D |
| A | 828 | 168 | 58 | 43 |
| B | 266 | 365 | 86 | 33 |
| C | 68 | 86 | 169 | 13 |
| D | 105 | 37 | 7 | 471 |
| Accuracy | (828 + 365 + 169 + 471) / 2803 = 65.39 % | | | |

| SVM Using Test Set Having Boolean Word Count | | | | |
|--|---|-----|----|-----|
| | A | B | C | D |
| A | 526 | 18 | 0 | 0 |
| B | 93 | 279 | 2 | 0 |
| C | 119 | 22 | 26 | 1 |
| D | 73 | 4 | 2 | 231 |
| Accuracy | (526 + 279 + 26 + 231) / 1396 = 76.07 % | | | |

| SVM Using 10-Fold Cross-Validation Having Boolean Word Count | | | | |
|--|--|-----|----|-----|
| | A | B | C | D |
| A | 1078 | 14 | 1 | 4 |
| B | 257 | 489 | 1 | 3 |
| C | 248 | 58 | 29 | 1 |
| D | 148 | 5 | 1 | 466 |
| Accuracy | (1078 + 489 + 29 + 466) / 2803 = 73.56 % | | | |

| SVM Using Test Set Having Boolean Word Count | | | | |
|--|---|-----|----|-----|
| | A | B | C | D |
| A | 515 | 23 | 4 | 2 |
| B | 89 | 267 | 8 | 10 |
| C | 64 | 19 | 84 | 1 |
| D | 66 | 4 | 7 | 233 |
| Accuracy | (526 + 279 + 26 + 231) / 1396 = 78.72 % | | | |

| SVM Using 10-Fold Cross-Validation Having Boolean Word Count | | | | |
|--|---|-----|-----|-----|
| | A | B | C | D |
| A | 1049 | 34 | 10 | 4 |
| B | 257 | 469 | 15 | 9 |
| C | 133 | 47 | 142 | 14 |
| D | 123 | 14 | 5 | 478 |
| Accuracy | (1049 + 469 + 142 + 478) / 2803 = 76.28 % | | | |

In my experiments, I wanted to compare how word count would affect both classifiers. I elected to convert first batch of files to vector based files with Boolean word count and we saw that on average accuracy was around 76.00%; however, when I converted files to vector based files with actual integer word count the average dropped to 70.95%. Nevertheless, the most accurate classifier was SVM using actual integer word count tested against test file and the accuracy was calculated to 78.72%.

Files used for classification are available under this link:

<https://www.dropbox.com/s/yw6guvubqrxgcec/Homework%20-%2002%20-%20Datasets.zip?dl=0>

Program code solution is available under this link:

<https://www.dropbox.com/s/pnv0ngppdyrb1u6/Homework%20-%2002%20-%20Code.zip?dl=0>

Addendum – Program used for conversion from txt format to arff format using C#

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.IO;
using System.Text;
using System.Threading.Tasks;

namespace CAP6776_HW2
{
    class Program
    {
        private static List<List<string>> tokens = new List<List<string>>();
        private static List<string> sentences = new List<string>();

        static void Main(string[] args)
        {
            CreateFilesARFF("webkb-train-stemmed.txt", "Train");
            CreateFilesARFF("webkb-test-stemmed.txt", "Test");
        }

        public static void CreateFilesARFF(string filename, string filetype)
        {
            List<string> sentences = ReadFile(filename);
            sentences = ConvertFile(sentences, filetype);
            WriteFile(filename.Substring(0, filename.Length - 4) + ".arff", sentences);
        }

        public static List<string> ReadFile(string filename)
        {
            List<string> result = new List<string>();
            var lines = File.ReadLines(filename);
            foreach (var line in lines)
            {
                result.Add(line);
            }
            return result;
        }

        public static void WriteFile(string filename, List<string> sentences)
        {
            File.WriteAllLines(filename, sentences);
        }

        public static List<string> ConvertFile(List<string> sentences, string type)
        {
            List<string> result = new List<string>();
            result.Add("@relation 'WebKB '" + type);
            result.Add("");
            result.Add("@attribute Text string");
            result.Add("@attribute class-att {student, faculty, project course}");
            result.Add("");
        }
    }
}
```

```
        result.Add("@data");
        result.Add("");
        foreach (var line in sentences)
        {
            int index = Search(line);
            string bodyClass = line.Substring(0, index);
            string bodyText = "\"" + line.Substring(index + 1, line.Length - index - 1) + "\"";
            result.Add(bodyText + "," + bodyClass);
        }
        return result;
    }

    public static int Search(string line)
    {
        for (int i = 0; i < line.Length; i++)
        {
            if (line[i] == '\\t') return i;
            {

            }
        }
        return -1;
    }
}
```