King Saud University
College of Computer and Information Sciences
Computer Science Department

CSC 569 SELECTED TOPICS IN AI

Assignment 3

The program description:

The program starts by reading the goldstd.csv file, then opens each link and reads the title and the announcement. With each time it retrieves a content it will tokenize it. After it reads all the contents it will remove the stopping words besides any word less than 3 letters, then apply stemming to them. By this step, the inverted index will be built as the previous assignment, so it will sort the index by IDF then it will write a CSV file where the columns are the attributes and the label, while the rows are the documents, and the values of attributes will be the documents frequencies.

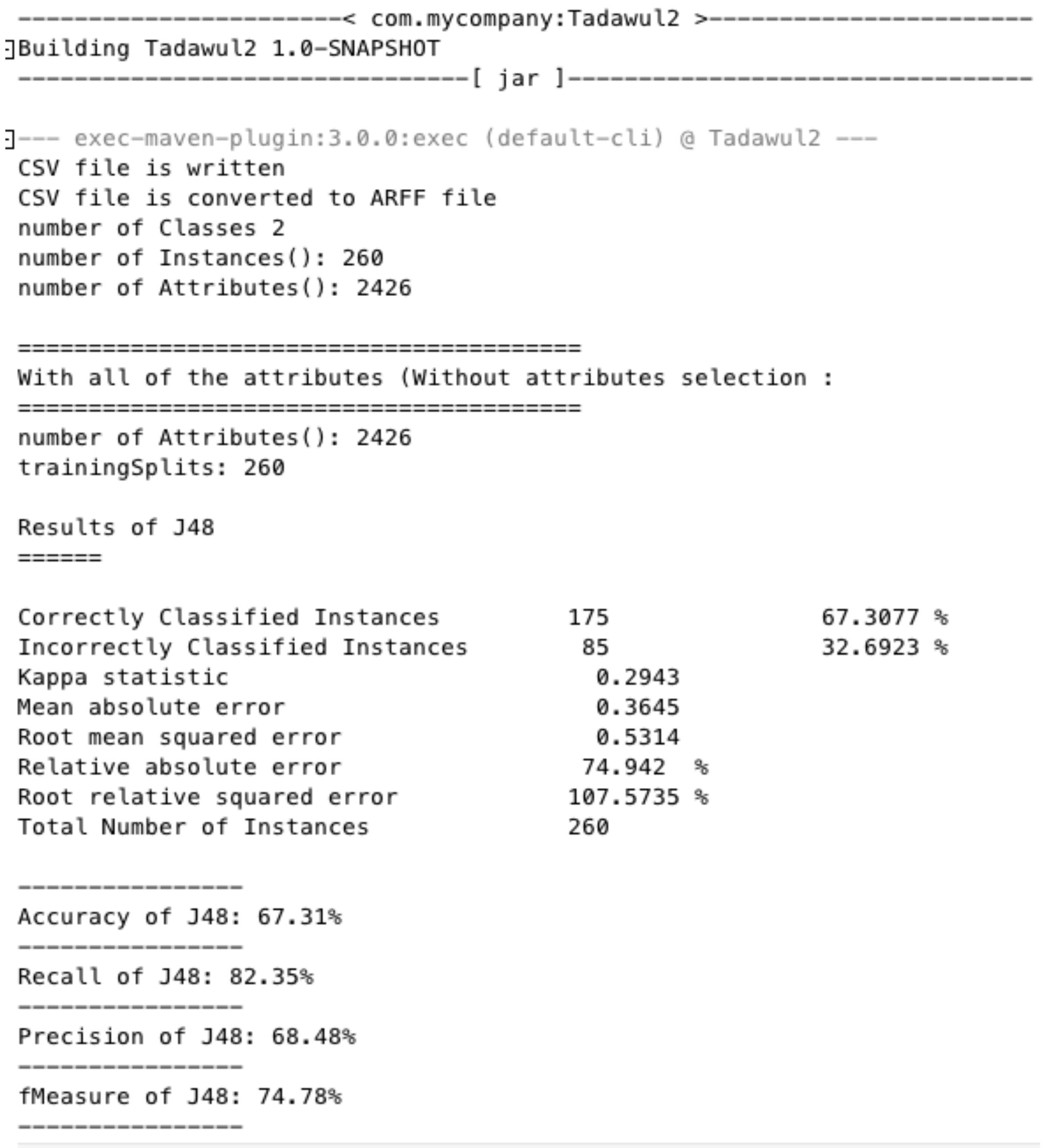
I tried two features (attributes) selection methods, the first is selecting the 50 topmost IDF of the attributes, second is using the information gain method in Weka to select the 50 attributes. Also, I tried to select 2000 attributes using the IDF method, then select 50 attributes from the 2000 attributes using the IG method.

The program will read the CSV file and convert it to an ARFF file, then apply the feature selection method, after that it will build two models (j-48 and SVM) with 260-fold cross-validation. Then it will display the evaluation metrics results, Accuracy, precision, recall, F measures.

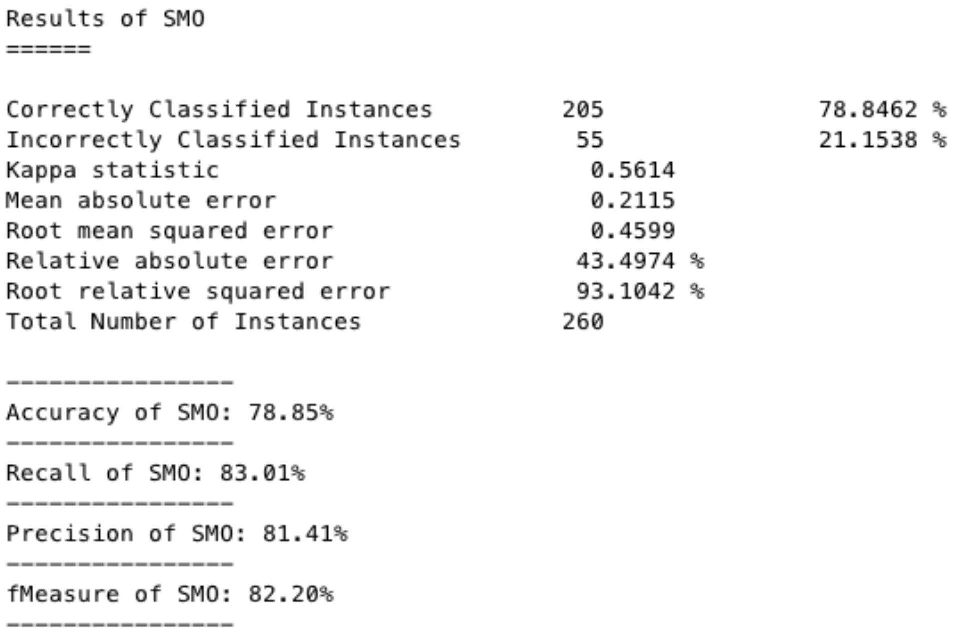
The results:

Attribute selection method	*Without attributes selection*		*IDF (50)*		*IG (50)*		*IDF (2000) then IG (50)*	
Model *Metric*	J48	SVM	J48	SVM	J48	SVM	J48	SVM
Accuracy	67.31%	78.85%	58.85%	58.85%	79.23%	86.92%	97.69%	98.46%
Recall	82.35%	83.01%	100.00%	100.00%	97.30%	94.59%	99.19%	99.59%
Precision	68.48%	81.41%	58.85%	58.85%	86.40%	94.59%	98.39%	98.79%
f-Measure	74.78%	82.20%	74.09%	74.09%	91.53%	94.59%	98.79%	99.19%

The output screen shots:



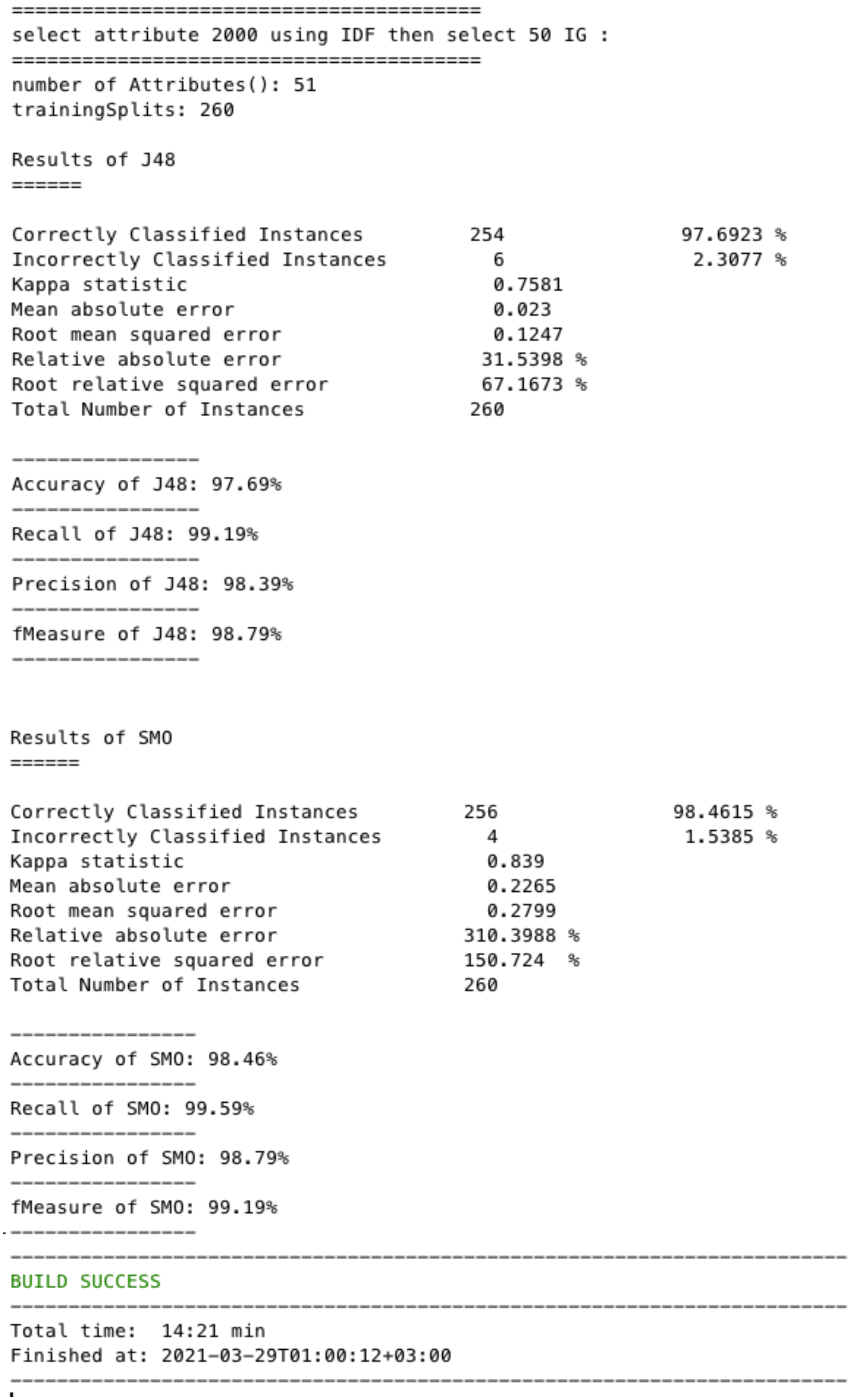
```
 -----------------------< com.mycompany:Tadawul2 >-----------------------
]Building Tadawul2 1.0-SNAPSHOT
 --------------------------------[ jar ]---------------------------------

]--- exec-maven-plugin:3.0.0:exec (default-cli) @ Tadawul2 ---
 CSV file is written
 CSV file is converted to ARFF file
 number of Classes 2
 number of Instances(): 260
 number of Attributes(): 2426

 ========================================
 With all of the attributes (Without attributes selection :
 ========================================
 number of Attributes(): 2426
 trainingSplits: 260

 Results of J48
 ======

 Correctly Classified Instances         175               67.3077 %
 Incorrectly Classified Instances        85               32.6923 %
 Kappa statistic                          0.2943
 Mean absolute error                      0.3645
 Root mean squared error                  0.5314
 Relative absolute error                 74.942  %
 Root relative squared error            107.5735 %
 Total Number of Instances              260

 ----------------
 Accuracy of J48: 67.31%
 ----------------
 Recall of J48: 82.35%
 ----------------
 Precision of J48: 68.48%
 ----------------
 fMeasure of J48: 74.78%
 ----------------
```



```
Results of SMO
======

Correctly Classified Instances         205               78.8462 %
Incorrectly Classified Instances        55               21.1538 %
Kappa statistic                          0.5614
Mean absolute error                      0.2115
Root mean squared error                  0.4599
Relative absolute error                 43.4974 %
Root relative squared error             93.1042 %
Total Number of Instances              260

----------------
Accuracy of SMO: 78.85%
----------------
Recall of SMO: 83.01%
----------------
Precision of SMO: 81.41%
----------------
fMeasure of SMO: 82.20%
----------------
```

```
========================================
select attribute 50 using IDF :
========================================
number of Attributes(): 51
trainingSplits: 260

Results of J48
======

Correctly Classified Instances         153               58.8462 %
Incorrectly Classified Instances       107               41.1538 %
Kappa statistic                          0
Mean absolute error                      0.4862
Root mean squared error                  0.494
Relative absolute error                 99.9783 %
Root relative squared error            100.0029 %
Total Number of Instances              260

----------------
Accuracy of J48: 58.85%
----------------
Recall of J48: 100.00%
----------------
Precision of J48: 58.85%
----------------
fMeasure of J48: 74.09%
----------------


Results of SMO
======

Correctly Classified Instances         153               58.8462 %
Incorrectly Classified Instances       107               41.1538 %
Kappa statistic                          0
Mean absolute error                      0.4115
Root mean squared error                  0.6415
Relative absolute error                 84.6221 %
Root relative squared error            129.8613 %
Total Number of Instances              260

----------------
Accuracy of SMO: 58.85%
----------------
Recall of SMO: 100.00%
----------------
Precision of SMO: 58.85%
----------------
fMeasure of SMO: 74.09%
----------------
```

```
========================================
select attribute 50 using IG :
========================================
number of Attributes(): 51
trainingSplits: 260

Results of J48
======

Correctly Classified Instances         206               79.2308 %
Incorrectly Classified Instances        54               20.7692 %
Kappa statistic                          0.7079
Mean absolute error                      0.044 
Root mean squared error                  0.1845
Relative absolute error                 30.1097 %
Root relative squared error             68.4178 %
Total Number of Instances              260     

----------------
Accuracy of J48: 79.23%
----------------
Recall of J48: 97.30%
----------------
Precision of J48: 86.40%
----------------
fMeasure of J48: 91.53%
----------------



Results of SMO
======

Correctly Classified Instances         226               86.9231 %
Incorrectly Classified Instances        34               13.0769 %
Kappa statistic                          0.8177
Mean absolute error                      0.1617
Root mean squared error                  0.2753
Relative absolute error                110.629  %
Root relative squared error            102.0784 %
Total Number of Instances              260     

----------------
Accuracy of SMO: 86.92%
----------------
Recall of SMO: 94.59%
----------------
Precision of SMO: 94.59%
----------------
fMeasure of SMO: 94.59%
----------------
```



```
========================================
select attribute 2000 using IDF then select 50 IG :
========================================
number of Attributes(): 51
trainingSplits: 260

Results of J48
======

Correctly Classified Instances         254               97.6923 %
Incorrectly Classified Instances         6                2.3077 %
Kappa statistic                          0.7581
Mean absolute error                      0.023
Root mean squared error                  0.1247
Relative absolute error                 31.5398 %
Root relative squared error             67.1673 %
Total Number of Instances              260

----------------
Accuracy of J48: 97.69%
----------------
Recall of J48: 99.19%
----------------
Precision of J48: 98.39%
----------------
fMeasure of J48: 98.79%
----------------


Results of SMO
======

Correctly Classified Instances         256               98.4615 %
Incorrectly Classified Instances         4                1.5385 %
Kappa statistic                          0.839
Mean absolute error                      0.2265
Root mean squared error                  0.2799
Relative absolute error                310.3988 %
Root relative squared error            150.724  %
Total Number of Instances              260

----------------
Accuracy of SMO: 98.46%
----------------
Recall of SMO: 99.59%
----------------
Precision of SMO: 98.79%
----------------
fMeasure of SMO: 99.19%
----------------
------------------------------------------------------------------------
BUILD SUCCESS
------------------------------------------------------------------------
Total time:  14:21 min
Finished at: 2021-03-29T01:00:12+03:00
------------------------------------------------------------------------
```

Shahd Taleb Alotaibi id: 438203473