CS 6375 – Machine Learning
Homework#2

Name: Revanth Segu

The program expects four arguments (< training_directory_path> < stop_words_file_path> < test_directory_path> <smoothing_feature>).

training_ directory _path : absolute path of Training files (in spam and ham directories)
stop_words_file_path : absolute path of stop words file(one word in each line)
test_ directory _path : absolute path of Test set file (in spam and ham directories)
smoothing_feature :{yes, no} yes – executes with smoothing feature

If any of the paths mentioned above in the same directory as class files then '.' can be given instead of absolute path.

The program can be executed by following the four steps given below:

1.javac Stemmer.java
2.javac TokenFrequency.java
3.javac ClassifyText.java
4.java ClassifyText . . . yes

In the above case training_files, stop_words_file, test_files are present in same directory so '.' is given instead of absolute path.

For Logistic Regression there are methods for printing Phat on each iteration and weights of token after iterations and a class variable "hardLimit" specifies the number of iterations to be run similarly neta and alpha are also can be configured

For Logistic Regression
Uncomment line //printPhats(); (to print Phats after each iteration)
Uncomment line - // printtokenWs(); (to print token weights)

Accuracies of Text Classifiers (with and without removing short list of stop words and smoothing feature):

Naïve Bayes – without smoothing
Without removing stop words
Accuracy of Ham : 91.95%
Accuracy of Spam : 98.46%
Total Accuracy : 93.72%

Removing stop words
Accuracy of Ham : 91.95%
Accuracy of Spam : 98.46%
Total Accuracy : 93.72%

Naïve Bayes – with smoothing
Without removing stop words
Accuracy of Ham : 95.11%

Accuracy of Spam : 99.23%
Total Accuracy : 96.23%

Removing stop words
Accuracy of Ham : 94.25%
Accuracy of Spam : 99.23%
Total Accuracy : 95.61%

Logistic Regression Classifier – without smoothing:
For Iteration- 100 alpha-0.001 Program takes approximately 4 to 8 minutes to execute
For Iteration- 10 alpha-0.1 Program takes approximately 1 to 2 minutes to execute

Logistic Regression – Without removing stop words:

# Iterations (down) Alpha→	0.1	0.01	0.001
10	28.66%	78.45%	76.78%
20	72.8%	87.45%	92.05%
50	72.8%	72.8%	92.05%
100	72.8%	72.8%	91.0%

Logistic Regression – Removing stop words:

# Iterations (down) Alpha→	0.1	0.01	0.001
10	25.52%	79.08%	84.52%
20	72.8%	91.0%	93.1%
50	72.8%	72.8%	94.35%
100	72.8%	72.8%	94.35%

Logistic Regression Classifier – with smoothing:
For Iteration- 100 alpha-0.001 Program takes approximately 4 to 8 minutes to execute
For Iteration- 10 alpha-0.1 Program takes approximately 1 to 2 minutes to execute

Logistic Regression – Without removing stop words:

# Iterations (down) Alpha→	0.1	0.01	0.001
10	75.1%	90.79%	89.75%
20	72.8%	89.96%	90.59%
50	72.8%	49.79%	91.21%
100	72.8%	72.8%	90.79%

Logistic Regression – Removing stop words:

# Iterations (down) Alpha→	0.1	0.01	0.001
10	91.84%	94.56%	94.56%
20	76.36%	94.77%	94.56%
50	72.8%	94.77%	94.56%
100	72.8%	72.8%	95.19%

Effect of filtering stop words:

Naïve Bayes classifier:- Removing stop words will reduce the accuracy of ham files as there are more stop words in ham files than spam files.

Logistic Regression classifier:- Removal of stop words will increase the accuracy as we don’t need to learn weights of those tokens and the express ability of classifier reduces which reduces overfitting.

Effect of Smoothing:

For smoothing porter stemmer have been to improve accuracy (porter stemmer stems the word and gives the root form of the word) other than that few other tokes are discarded in the text while training which also significantly improved accuracy. Please see constructor of class ClassifyText.java file where the list of words to be discarded are added to list.

Naïve Bayes classifier:- Smoothing feature increases accuracy as the un important tokens are removed and stemming maps all tenses of a verb to its root form (like propagate, propagation, propagated are mapped to the same root word).

Logistic Regression classifier:- Smoothing feature increases accuracy as the un important tokens are removed and stemming maps all tenses of a verb to its root form (like propagate, propagation, propagated are mapped to the same root word).

Accuracy of Classifier for each individual spam and ham classes will be printed if the below shown code is uncommented

```
/*percentage = (double)(correctHamClassify)/(double)(testHam);

percentage *= 100;
percentage = roundTwoDecimals(percentage);

System.out.println("Accuracy of Ham for Naive Bayes Classifier: "+percentage + "%");

percentage = (double)(correctSpamClassify)/(double)(testSpam);

percentage *= 100;
percentage = roundTwoDecimals(percentage);



System.out.println("Accuracy of Spam for Naive Bayes Classifier: "+percentage +
"%");*/
```