CS 6375 – Machine Learning
Homework#3

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 BuildPerceptron.java
4.java BuildPerceptron . . . yes

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

There are methods for printing Perceptron utputs on each iteration and weights of token after iterations and a class variable "hardLimit" specifies the number of iterations to be run similarly neta can be configured

For Logistic Regression
Uncomment line // printPouts(); (to print Perceptron outputs after each iteration)
Uncomment line - // printtokenWs(); (to print token weights)

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

Results of Hw2 Dataset:

Without smoothing:

For Iteration- 100 neta-0.001 Program takes approximately 90 seconds to execute
For Iteration- 10 neta-0.1 Program takes approximately 20 seconds to execute

Without removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	87.5%	89.75%	89.75%	89.75%	89.75%
20	91.21%	91.21%	91.21%	91.21%	91.0%
50	91.21%	91.21%	91.21%	91.21%	91.21%
100	91.21%	91.21%	91.21%	91.21%	91.21%

Removing stop words:

# Iterations	0.1	0.05	0.01	0.005	0.001

(down) Neta→					
10	91.42%	91.42%	91.42%	91.42%	91.42%
20	92.05%	92.05%	92.05%	92.05%	92.05%
50	92.05%	92.05%	92.05%	92.05%	92.05%
100	92.05%	92.05%	92.05%	92.05%	92.05%

Without smoothing:
For Iteration- 100 neta-0.001 Program takes approximately 90 seconds to execute
For Iteration- 10 neta-0.1 Program takes approximately 20 seconds to execute

Without removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	78.87%	78.87%	78.87%	78.87%	78.87%
20	83.68%	83.68%	83.68%	83.68%	83.68%
50	91.63%	91.63%	91.63%	91.21%	91.21%
100	91.42%	91.42%	91.42%	91.42%	91.42%

Removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	73.85%	73.85%	73.85%	73.85%	73.85%
20	83.47%	83.47%	83.47%	83.47%	83.47%
50	91.63%	91.63%	91.63%	91.63%	91.63%
100	92.47%	92.47%	92.47%	92.47%	92.47%

Results of Enron1 Dataset:
Without smoothing:
For Iteration- 100 neta-0.001 Program takes approximately 90 seconds to execute
For Iteration- 10 neta-0.1 Program takes approximately 20 seconds to execute

Without removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	92.54%	92.54%	92.54%	92.54%	92.54%
20	93.2%	93.2%	93.2%	93.2%	92.76%
50	92.76%	92.76%	92.76%	92.76%	92.76%
100	92.76%	92.76%	92.76%	92.76%	92.76%

Removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	90.13%	90.13%	90.13%	90.13%	90.13%
20	92.35%	90.35%	90.35%	90.57%	90.35%
50	92.11%	92.11%	92.11%	92.54%	91.45%
100	92.32%	92.32%	92.32%	92.54%	92.11%

Without smoothing:
For Iteration- 100 neta-0.001 Program takes approximately 90 seconds to execute
For Iteration- 10 neta-0.1 Program takes approximately 20 seconds to execute

Without removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	77.41%	77.41%	77.41%	77.41%	77.41%
20	88.38%	88.38%	88.38%	88.38%	88.38%
50	92.54%	92.54%	92.54%	92.54%	92.11%
100	91.89%	91.89%	91.89%	91.89%	91.67%

Removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	73.68%	73.68%	73.68%	73.68%	73.68%
20	87.27%	87.27%	87.27%	87.27%	87.94%
50	94.52%	94.52%	94.52%	94.52%	93.64%
100	94.3%	94.3%	94.3%	94.3%	94.52%

Results of Enron4 Dataset:
Without smoothing:
For Iteration- 100 neta-0.001 Program takes approximately 250 seconds to execute
For Iteration- 10 neta-0.1 Program takes approximately 40 seconds to execute

Without removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	93.0%	93.0%	94.11%	93.0%	93.0%
20	93.37%	93.37%	93.37%	93.37%	93.37%
50	93.74%	93.74%	93.74%	93.74%	93.74%
100	94.11%	94.11%	94.11%	94.29%	94.11%

Removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	94.66%	94.66%	94.66%	94.66%	94.66%
20	94.66%	94.66%	94.66%	94.66%	94.66%
50	94.66%	94.66%	94.66%	94.66%	94.66%
100	94.66%	94.66%	94.66%	94.66%	94.66%

Without smoothing:
For Iteration- 100 neta-0.001 Program takes approximately 250 seconds to execute
For Iteration- 10 neta-0.1 Program takes approximately 40 seconds to execute

Without removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	86.37%	86.37%	86.37%	86.37%	86.37%

20	89.69%	89.69%	89.69%	89.69%	89.69%
50	95.21%	95.21%	95.21%	95.21%	95.21%
100	95.4%	95.4%	95.4%	95.4%	95.4%

Removing stop words:

# Iterations (down) Neta→	0.1	0.05	0.01	0.005	0.001
10	81.58%	81.58%	81.58%	81.58%	81.58%
20	93.74%	93.74%	93.74%	93.74%	93.74%
50	96.13%	96.13%	96.13%	96.13%	96.13%
100	96.32%	96.32%	96.32%	96.32%	96.32%

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 Perceptron Classifier: "+percentage + "%");

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

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


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

SVM Results

Data set Hw#2

Linear Kernel:

=== Summary ===

Correctly Classified Instances	421	92.3246 %
Incorrectly Classified Instances	35	7.6754 %
Kappa statistic	0.8271	
Mean absolute error	0.0768	
Root mean squared error	0.277	

Relative absolute error 17.4938 %

Root relative squared error 59.1695 %

Total Number of Instances 456

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.932	0.095	0.953	0.932	0.943	0.919	1
	0.905	0.068	0.865	0.905	0.884	0.919	-1
Weighted Avg.	0.923	0.086	0.925	0.923	0.924	0.919	

=== Confusion Matrix ===

a	b	<-- classified as
287	21	a = 1
14	134	b = -1

Polynomial Kernel:

=== Summary ===

Correctly Classified Instances 309 67.7632 %

Incorrectly Classified Instances 147 32.2368 %

Kappa statistic 0.0091

Mean absolute error 0.3224

Root mean squared error 0.5678

Relative absolute error 73.474 %

Root relative squared error 121.2613 %

Total Number of Instances 456

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0.993	0.677	1	0.807	0.503	1
	0.007	0	1	0.007	0.013	0.503	-1
Weighted Avg.	0.678	0.671	0.782	0.678	0.55	0.503	

=== Confusion Matrix ===

a	b	<-- classified as
308	0	a = 1
147	1	b = -1

Sigmoid Kernel:

=== Summary ===

Correctly Classified Instances	310	67.9825 %
Incorrectly Classified Instances	146	32.0175 %
Kappa statistic	0.0182	
Mean absolute error	0.3202	
Root mean squared error	0.5658	
Relative absolute error	72.9742 %	
Root relative squared error	120.8482 %	
Total Number of Instances	456	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0.986	0.678	1	0.808	0.507	1
	0.014	0	1	0.014	0.027	0.507	-1
Weighted Avg.	0.68	0.666	0.783	0.68	0.555	0.507	

=== Confusion Matrix ===

a	b	<-- classified as
308	0	a = 1
146	2	b = -1

Data set Enron1

Linear Kernel:

=== Summary ===

Correctly Classified Instances	421	92.3246 %
Incorrectly Classified Instances	35	7.6754 %
Kappa statistic	0.8271	
Mean absolute error	0.0768	
Root mean squared error	0.277	
Relative absolute error	17.4938 %	
Root relative squared error	59.1695 %	
Total Number of Instances	456	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.932	0.095	0.953	0.932	0.943	0.919	1
	0.905	0.068	0.865	0.905	0.884	0.919	-1
Weighted Avg.	0.923	0.086	0.925	0.923	0.924	0.919	

=== Confusion Matrix ===

a	b	<-- classified as
287	21	a = 1
14	134	b = -1

Polynomial Kernel:

=== Summary ===

Correctly Classified Instances	309	67.7632 %
Incorrectly Classified Instances	147	32.2368 %
Kappa statistic	0.0091	
Mean absolute error	0.3224	
Root mean squared error	0.5678	
Relative absolute error	73.4676 %	
Root relative squared error	121.2608 %	
Total Number of Instances	456	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0.993	0.677	1	0.807	0.503	1

	0.007	0	1	0.007	0.013	0.503	-1
Weighted Avg.	0.678	0.671	0.782	0.678	0.55	0.503	

=== Confusion Matrix ===

a	b	<-- classified as
308	0	a = 1
147	1	b = -1

Sigmoid Kernel:

=== Summary ===

Correctly Classified Instances	310	67.9825 %
Incorrectly Classified Instances	146	32.0175 %
Kappa statistic	0.0182	
Mean absolute error	0.3202	
Root mean squared error	0.5658	
Relative absolute error	72.9678 %	
Root relative squared error	120.8476 %	
Total Number of Instances	456	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.986	0.678	1	0.808	0.507	1
0.014	0	1	0.014	0.027	0.507	-1

Weighted Avg. 0.68 0.666 0.783 0.68 0.555 0.507

=== Confusion Matrix ===

a	b	<-- classified as
308	0	a = 1
146	2	b = -1

Data Set Enron4:

Linear Kernel:

=== Summary ===

Correctly Classified Instances	515	94.8435 %
Incorrectly Classified Instances	28	5.1565 %
Kappa statistic	0.87	
Mean absolute error	0.0516	
Root mean squared error	0.2271	
Relative absolute error	12.6777 %	
Root relative squared error	50.3767 %	
Total Number of Instances	543	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.87	0.021	0.944	0.87	0.905	0.925	1
0.979	0.13	0.95	0.979	0.965	0.925	-1
Weighted Avg. 0.948	0.099	0.948	0.948	0.948	0.925	

=== Confusion Matrix ===

a b <-- classified as

134 20 | a = 1

8 381 | b = -1

Polynomial Kernel:

== Summary ===

Correctly Classified Instances	389	71.639 %
Incorrectly Classified Instances	154	28.361 %
Kappa statistic	0	
Mean absolute error	0.2836	
Root mean squared error	0.5326	
Relative absolute error	69.6767 %	
Root relative squared error	118.1467 %	
Total Number of Instances	543	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0	0	0	0	0	0.5	1
	1	1	0.716	1	0.835	0.5	-1
Weighted Avg.	0.716	0.716	0.513	0.716	0.598	0.5	

=== Confusion Matrix ===

a b <-- classified as

0 154 | a = 1

0 389 | b = -1

Sigmoid Kernel:

=== Summary ===

Correctly Classified Instances	406	74.7698 %
Incorrectly Classified Instances	137	25.2302 %
Kappa statistic	0.151	
Mean absolute error	0.2523	
Root mean squared error	0.5023	
Relative absolute error	62.0236 %	
Root relative squared error	111.4342 %	
Total Number of Instances	543	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.11	0	1	0.11	0.199	0.555	1
	1	0.89	0.74	1	0.85	0.555	-1
Weighted Avg.	0.748	0.637	0.813	0.748	0.666	0.555	

=== Confusion Matrix ===

a	b	<-- classified as
17	137	a = 1
0	389	b = -1

Neural Networks Results

Data set Hw#2

Hidden Layers:1

=== Summary ===

Correctly Classified Instances	421	92.3246 %

Incorrectly Classified Instances 35 7.6754 %

Kappa statistic 0.8271

Mean absolute error 0.0768

Root mean squared error 0.277

Relative absolute error 17.4938 %

Root relative squared error 59.1695 %

Total Number of Instances 456

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.932	0.095	0.953	0.932	0.943	0.919	1
0.905	0.068	0.865	0.905	0.884	0.919	-1
Weighted Avg. 0.923	0.086	0.925	0.923	0.924	0.919	

=== Confusion Matrix ===

a	b	<-- classified as
287	21	a = 1
14	134	b = -1

Hidden Layers 2

=== Summary ===

Correctly Classified Instances 309 67.7632 %

Incorrectly Classified Instances 147 32.2368 %

Kappa statistic 0.0091

Mean absolute error 0.3224

Root mean squared error 0.5678

Relative absolute error 73.474 %

Root relative squared error 121.2613 %

Total Number of Instances 456

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0.993	0.677	1	0.807	0.503	1
	0.007	0	1	0.007	0.013	0.503	-1
Weighted Avg.	0.678	0.671	0.782	0.678	0.55	0.503	

=== Confusion Matrix ===

a	b	<-- classified as
308	0	a = 1
147	1	b = -1

Hidden Layers 3

=== Summary ===

Correctly Classified Instances 310 67.9825 %

Incorrectly Classified Instances 146 32.0175 %

Kappa statistic 0.0182

Mean absolute error 0.3202

Root mean squared error 0.5658

Relative absolute error 72.9742 %

Root relative squared error 120.8482 %

Total Number of Instances 456

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0.986	0.678	1	0.808	0.507	1
	0.014	0	1	0.014	0.027	0.507	-1
Weighted Avg.	0.68	0.666	0.783	0.68	0.555	0.507	

=== Confusion Matrix ===

a	b	<-- classified as
308	0	a = 1
146	2	b = -1

Data set Enron1

Hidden Layers 1

=== Summary ===

Correctly Classified Instances 421 92.3246 %

Incorrectly Classified Instances 35 7.6754 %

Kappa statistic 0.8271

Mean absolute error 0.0768

Root mean squared error 0.277

Relative absolute error 17.4938 %

Root relative squared error 59.1695 %

Total Number of Instances 456

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.932	0.095	0.953	0.932	0.943	0.919	1
	0.905	0.068	0.865	0.905	0.884	0.919	-1
Weighted Avg.	0.923	0.086	0.925	0.923	0.924	0.919	

=== Confusion Matrix ===

a	b	<-- classified as
287	21	a = 1
14	134	b = -1

Hidden Layers 2

=== Summary ===

Correctly Classified Instances	309	67.7632 %
Incorrectly Classified Instances	147	32.2368 %
Kappa statistic	0.0091	
Mean absolute error	0.3224	
Root mean squared error	0.5678	
Relative absolute error	73.4676 %	
Root relative squared error	121.2608 %	
Total Number of Instances	456	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0.993	0.677	1	0.807	0.503	1
	0.007	0	1	0.007	0.013	0.503	-1
Weighted Avg.	0.678	0.671	0.782	0.678	0.55	0.503	

=== Confusion Matrix ===

a	b	<-- classified as
308	0	a = 1
147	1	b = -1

Hidden Layers 3

=== Summary ===

Correctly Classified Instances	310	67.9825 %
Incorrectly Classified Instances	146	32.0175 %
Kappa statistic	0.0182	
Mean absolute error	0.3202	
Root mean squared error	0.5658	
Relative absolute error	72.9678 %	
Root relative squared error	120.8476 %	
Total Number of Instances	456	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0.986	0.678	1	0.808	0.507	1
	0.014	0	1	0.014	0.027	0.507	-1
Weighted Avg.	0.68	0.666	0.783	0.68	0.555	0.507	

=== Confusion Matrix ===

a	b	<-- classified as
308	0	a = 1
146	2	b = -1

Data Set Enron4:

Hidden Layers 1

=== Summary ===

Correctly Classified Instances	515	94.8435 %
Incorrectly Classified Instances	28	5.1565 %
Kappa statistic	0.87	
Mean absolute error	0.0516	
Root mean squared error	0.2271	
Relative absolute error	12.6777 %	
Root relative squared error	50.3767 %	
Total Number of Instances	543	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.87	0.021	0.944	0.87	0.905	0.925	1

0.979 0.13 0.95 0.979 0.965 0.925 -1

Weighted Avg. 0.948 0.099 0.948 0.948 0.948 0.925

=== Confusion Matrix ===

a b <-- classified as

134 20 | a = 1

8 381 | b = -1

Hidden Layers 2

== Summary ===

Correctly Classified Instances 389 71.639 %

Incorrectly Classified Instances 154 28.361 %

Kappa statistic 0

Mean absolute error 0.2836

Root mean squared error 0.5326

Relative absolute error 69.6767 %

Root relative squared error 118.1467 %

Total Number of Instances 543

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0	0	0	0	0.5	1
1	1	0.716	1	0.835	0.5	-1
Weighted Avg.	0.716	0.716	0.513	0.716	0.598	0.5

=== Confusion Matrix ===

a b <-- classified as

0 154 | a = 1

0 389 | b = -1

Hidden Layers 3

=== Summary ===

Correctly Classified Instances	406	74.7698 %
Incorrectly Classified Instances	137	25.2302 %
Kappa statistic	0.151	
Mean absolute error	0.2523	
Root mean squared error	0.5023	
Relative absolute error	62.0236 %	
Root relative squared error	111.4342 %	
Total Number of Instances	543	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.11	0	1	0.11	0.199	0.555	1
	1	0.89	0.74	1	0.85	0.555	-1
Weighted Avg.	0.748	0.637	0.813	0.748	0.666	0.555	

=== Confusion Matrix ===

a b <-- classified as

17 137 | a = 1

0 389 | b = -1