CSE 5525: HOMEWORK 2

TEXT CLASSIFICATION

-By Shiuli Das, 500043947

OBJECTIVE: Implement naïve bayes and perceptron algorithm for text classification. Train models on dataset of 25000 positive and negative IMDb movie reviews and report prediction accuracy on a test set.

RESULTS:

1. Naïve Bayes

The following table gives the accuracy obtained for different values of smoothing parameter ALPHA.

ALPHA	Accuracy (%)
0.1	82.32
0.5	82.356
1.0	82.408
2.5	82.628
5.0	82.832
7.0	82.936
8.0	83.044
9.0	83.072
10.0	83.112
20.0	83.204
30.0	83.184
100.0	83.24
120.0	83.264
125.0	**83.268**
150.0	83.216
200.0	83.172
500.0	83.14
1000.0	83.124

Obtained best accuracy of 83.268% with smoothing parameter ALPHA = 125

2. Perceptron

The following table gives the accuracy of the perceptron classification algorithm when it is run for varied number of iterations.

N_ITERATIONS	Accuracy (%)	
		Averaged Parameters
1	75.408	84.068
10	81.248	87.208
30	83.144	**87.22**
40	84.976	87.144
50	81.376	87.072
100	85.992	86.936
200	86.116	86.56
1000	**86.404**	86.22
2000	86.116	86.18

Accuracy first increases with increasing number of iterations as the weights get updated but then starts reducing on the test set since the weights get over-fitted to the training dataset. There are slight deviations from this general trend since the random initial weight vector chosen has some effect as well.

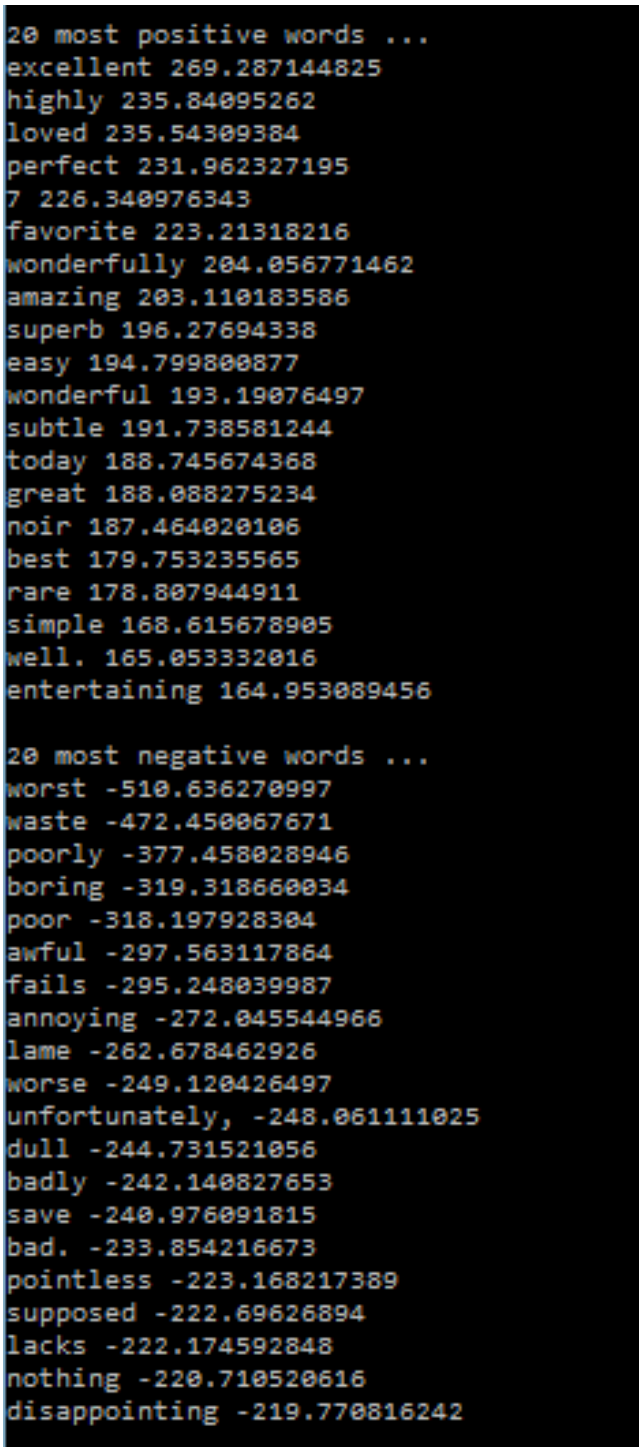
It is observed that using parameter averaging, we get better accuracy for lower number of iterations.

The program was run again, this time, the data was shuffled after each iteration.

N_ITERATIONS	Accuracy (%)	
		Averaged Parameters
1	75.408	84.068
5	85.776	86.86
10	**86.328**	**87.46**
30	86.068	87.364
40	85.58	87.288
50	86.068	87.236
100	86.272	86.832
200	86.272	86.612

We see that when the data is shuffled after each iteration, the classification converges faster.

Best accuracy of 87.46% obtained at N_ITERATIONS = 10, (using averaged parameters and shuffling data after each iteration)



CONCLUSION:

The naïve bayes and perceptron classification algorithm for text classification on IMDb movie reviews as positive and negative. Accuracy results have been presented for various hyper-parameter values.