CS 6375: Machine Learning
Assignment #3

Prof. Anjum Chida

Name: Akash Biswal

NetID: axb200166

Naive Bayes:

Metrics for Text classification using Multinomial Naive Bayes algorithm:

Total Test files (478) – Ham files (348) + Spam files (130)

Accuracy for classifying as “Ham”

With stop words included	After discarding stop words
89.94%	89.94%

Accuracy for classifying as “Spam”

With stop words included	After discarding stop words
90.77%	87.69%

Overall Accuracy for classifying the email correctly with stop words: 90.17%
Overall Accuracy for classifying the email correctly without stop words: 89.33%

Inference:
The extra stop words act as added features and therefore removing them decreases the accuracy of the algorithm.

Logistic Regression:

Metrics for Text classification using Logistic Regression algorithm:

Learning Rate is fixed as 0.001 throughout

For, Lamda = 0.01 and Iterations = 50
Accuracy on Ham: 92.24137931034483
Accuracy on Spam: 7.6923076923076925
Overall Accuracy : 69.24686192468619

Issues Faced:

- Weight Matrix being updated with underflowing values returned from the sigmoid function
- Due to this most files are being classified as spam