Edward Eastwood

Comp 379: HW4 Write Up

The purpose of this assignment was to create a Naive Bayes Classifier in Python. My implementation was about 50% accurate at best. I evaluated this by computing the number of correctly classified samples in proportion to the total number in the test data. I used a standard Naive Bayes implementation except this version was tailored to the data. Because of the two different data sets being given as massive text files, I used two different data structures for each. Though this is not the most memory efficient, it made the pre-processing go much smoother.

For pre-processing, I removed stop words, single characters and isolated special characters. I split the data into training and test sets. I had originally done a development set but it improved the accuracy to have more test data. This is because it allowed for a larger probability that the words in the test data were contained in the training data.