CSE 408 Project 1: Text Classification and Sentimental Analysis Report

For the first part of this project – the text classification portion – the algorithms did not work particularly accurately. This is primarily because they did not take into account the word "not." For instance, the sentiment analysis algorithms worked well when the positive reviews used a lot of positive words such as "happy" and "good," but did not work when there were phrases like "not bad" which a human would regard as a positive review, but the algorithm would classify it as a negative review. The text classification also had its faults again for the same reason. The text classification had about a 55-65% chance of identifying the review correctly, which is still more accurate than not accurate but because many positive and negative reviews are similar given the algorithm would consider phrases "good" and "not good" very similarly when a human would understand they have very different meanings.

In the KNN part, when using the sum of squared distances, the highest levels of accuracy came when K = 2 and K = 3. Both K = 2 and K = 3, yielded an accuracy of 60.53%. When using the angle between vectors method of calculating distance, the best value of K to return a high level of accuracy was in the range of 4 to 6. For K = 4, the accuracy was 65.79%; for K = 5, the accuracy was 68.42%; for K = 6 the accuracy was 65.79%. When using the number of words in common metric for measuring distance between documents, the highest levels of accuracy were provided for K = 6 and K = 7. K = 6 resulted an accuracy of 65.79% and K = 7 returned an accuracy of 63.16%. These results were found by running the algorithms for each value of K from 1 to 37. It makes sense that if K is too small it would not result in very high accuracy, given that there is not enough data to give an opinion given that if the two closest neighbors to a positive review might be negative, but the next 5 nearest reviews would be positive. It also makes sense that as the value of K approaches the whole data set, the accuracy falls again, given that there is an equal number of positive and negative reviews in the sample.

As can be seen in the above paragraph, the highest accuracy was found in the angle between vectors method of measuring distance and a K-Nearest Neighbors value of 5 nearest neighbors. Therefore, the angle between vectors is empirically the best way to classify the documents. However, my initial reaction would be to think that the number of words in common would result in the most accurate assessment, given that I would think a lot of positive reviews would use the same words, and the negative words would use the same reviews.

For the sentiment Analysis, the highest rated review that was in the negative folder was 04.txt and it was given a sentimental analysis score of about 5.51. This is because, the review uses a lot of positive words, and then describes why they did not like the application. For instance, the review said "the game are somewhat pleasant … But once playing the game, that is where my gripes start to come in." The algorithm rates the word "pleasant" (0.625) as positive but there is no place in the algorithm to get the "but." The lowest rated review that was in the positive folder was 01.txt which had the score of roughly -0.99. This is because the author uses words like "never" (-0.75) and "not" (-0.75) and "phone" (-0.625) but does not use many positive words even though the overall review has a positive tone that a human but not this base-level algorithm can pick up on.