Robert Nelson

CIS415

Assignment 4

My method as I was searching for the best accuracy was by reading through the provided articles and testing each variable, between the Boolean parameters and the string parameters. Essentially, I worked my way through all three lines of code and tested various settings to see whether the accuracy increased or decreased. For example, with the CountVectorizer, I decided to change stop\_words to ‘english’ because I figured that it would more accurately classify my data while set to English. I was surprised that lowercase set to False was more helpful for my data. I think this is because some movie titles can be capitalized incorrectly. Next, with TFIDF I also sorted through the Boolean parameters to find the optimal configuration. I found that use\_idf and smooth\_idf should both be True. I tested sublinear\_idf but it did not change my results significantly. For MultinomialNB, I tested alpha somewhat but found that 2 was a good float for this procedure. Again, I tested the Boolean parameters and with all of my manipulations, I arrived at 0.85 accuracy!



