Lab3 Report

# Lab3\_1

We initially did not understand the problem, but used the code from piazza to complete the code. We also used code snippets NLTK Chapter 6 because they were easy to understand and the only source of information.

Our solution takes Synsets (synonyms that share a common meaning) to train the Naive Bayes classifier. The Train set is ¾ of the total length of the corpus and the test set is ¼ of the corpus. The result differs every time because of the shuffle/code in line 14. The results are between 0.78 and 0.84, which is pretty high. This means that the classifier has an accuracy rate higher than 80% on average.

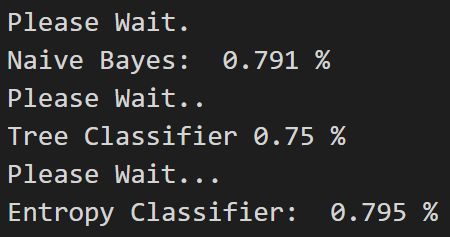
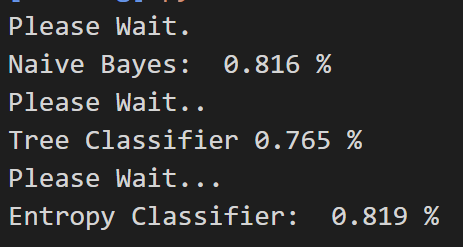
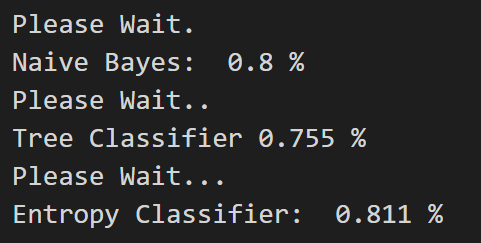




# Lab3\_2

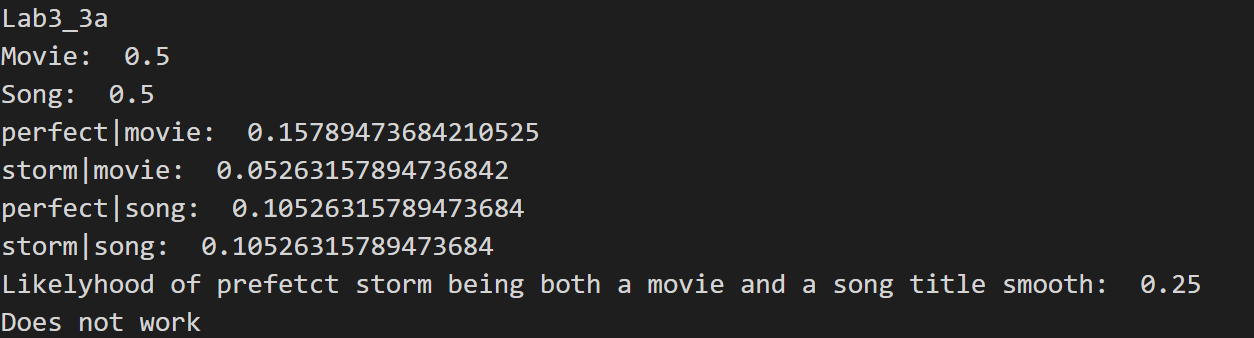
Here, we used the same training-set and test-set to see the similarity and difference between these three classifiers.

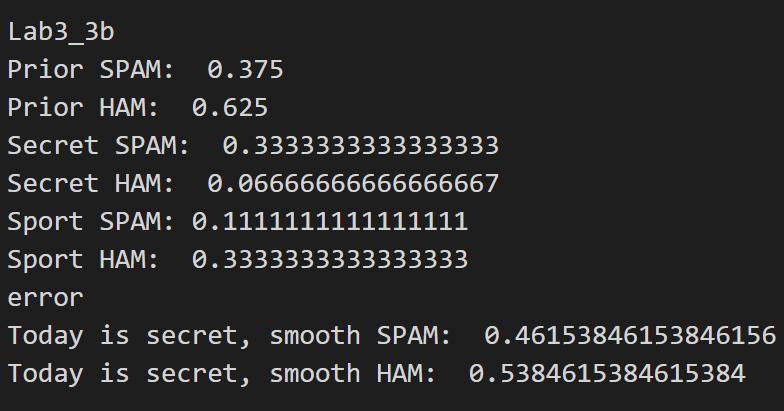
Used last letter, first letter, last two letters and last three letters in order to make the classification more specific. The result shows that Entropy Classifier and Naïve Bayes are the best and have similar results, while the tree classifier performs worse than the others. Overall, all three classifiers score high and the accuracy is between 75-80%.



# Lab3\_3

We don’t understand why the exception handles in line 124 and 146. 3a,1 gives 50% probability of the text being a movie or a song title.





# Lab3\_4

We could not find a solution for this task.

What we tried to do was make a corpus from tweets from two different accounts which we managed to do. Then we made a class PreProcessTweets which cleans up the tweets, removes the stopwords and unnecessary characters and the usernames and hashtags. This might be counterproductive but it’s a way to remove the special characters in the tweets. We then built a vocabulary, a list of the different words used per corpus and the frequency. We tried to solve the task similarly to the other exc1,2 with extract features, but we could not move forward from this point.

Another approach we did was to vectorize our corpus which we managed to do, but did not understand what to do next. Without any ideas on how to move forward we did not manage to complete it.