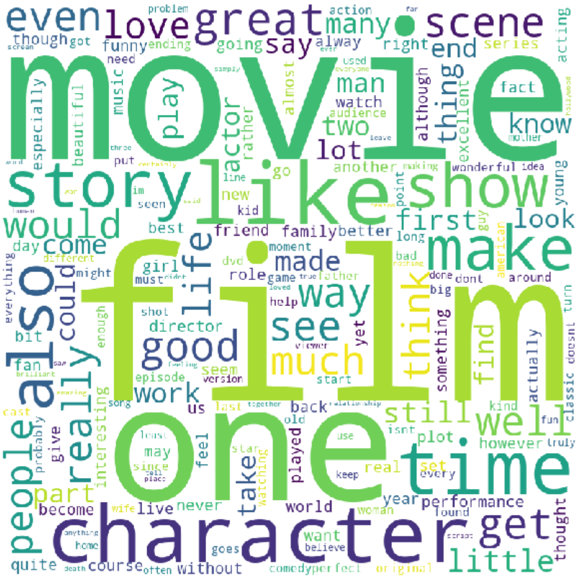
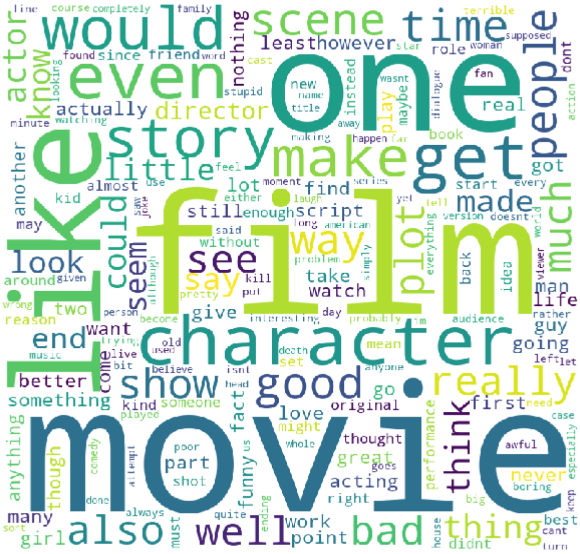
Naive Bayes

Assignment 5 report



Word cloud for the positive feedback words



Word cloud for the positive feedback words

One of the import parts of applying a machine learning model on text data is cleaning the data. For the purpose of the assignment the data was cleaned using removing all the non-alphabets and then removing all the stop words. In part b PorterStemmer was used to stem the words too. Stemming wasn't applied to words in part c because it was felt that stemming can make different adjectives lose their meaning and adjectives are very important when deciding the meaning of a sentence. As can be seen from the word cloud, the positive and negative words are almost similar with respect to the high frequency words. For the part a, the algorithm was simple. Count the frequency of words in positive and negative feedback and calculate probabilities and at the time of prediction, take the logarithmic sum of the probabilities using Laplace smoothening and then predict the label by comparing the probability value for positive and negative. Part b was same just with added stemming and stop words removal. The accuracy achieved on test set was 83.49% whereas accuracy on test set in part b reached 84.02%. For the part c, I tried different methods such as IDF-TF, inverted document frequency and term frequency but the accuracy only dropped. With IDF-TF, an accuracy of 70% was achieved on the training data. Then bigram was tried. Bigram when used along with bag of words gave an accuracy of 84% and when used alone gave an accuracy of 50.3% both on training data and therefore, I stopped experimenting with those. Increasing the probabilities of words that are adjectives may also help to improve the performance of the model. In part c a maximum accuracy of 84.63 was achieved on the test data.