NAÏVE BAYES ASSIGNMNET

A dataset for sentiments of movie review is available at http://ai.stanford.edu/~amaas/data/ sentiment/. The dataset contains movie reviews in natural language and their sentiment whether they are positive[1] or negative[-1]. We have processed this large dataset into three train sets [small, medium and large] and one test set. Use train-small dataset for training in sub-question a-f. Use all three train sets for question part g. There is a single test set, which should be used for all sub-questions. The dataset contains processed documents, the last column in each line denotes the label [1,-1], all other columns of vector represent encoded data [14666 columns]. These are large files and may take about 5 mins to load in matlab running on a desktop/laptop. Each index on the vector corresponds to the count of a vocabulary word [one-hot-vector encoding], imdb vocab.csv file contains the word for index i at line i (use this information for sub-question d).
Code for accuracy prediction is included in the code folder (Classification Accuracy.m). You can use 'confusionmat' function for computing confusion matrix in matlab (sklearn.metrics.confusion matrix in python).

Task:

Implement a naive bayes classifier for document classification for sentiment analysis dataset.
Please report train and test accuracies. Also provide the confusion matrix for test results.