Name - Sameer Vivek Pande
Entry Number - 2017CS10371

Naive Bayes – Part (a),(b)

We define feature vectors in which each feature represents a word in our dictionary. Our dictionary will contain all the words that were encountered in the training part. In part (b) we remove stopwords and use output of stemming in the dictionary.

We follow the Formula for naive bayes given in the *http://cs229.stanford.edu/notes/cs229-notes2.pdf* (which was also used in class).

Part (c)

Model 1 – Using port stemming, removal of stopwords, using bi-grams along with unigrams

Features - [unigrams]

Train accuracy = 92.73 %
Test accuracy = 85.14 %

Model 2 – Without port stemming or removal of stopwords. Using bi-grams along with uningrams

Features – [unigrams + bigrams]

Train accuracy = 94.2 %
Test accuracy = 85.09%

Model 3 – Without port stemming or removal of stopwords. Using only unigrams

Features – [unigrams]

Train accuracy = 94.8 %
Test accuracy = 85.0 %

Model 4 - Without port stemming or removal of stopwords, Using only bigrams

Features - [bigrams]

Train accurcacy = 50.7 %
Test accuracy = not tested due to very poor train accuracy

Model 5 – Using portstemming, removal of stopwords, using only unigrams

Features - [unigrams]
Train accuracy = 92.93 %
Test accuracy = 85.13 %

Model number 1 which uses bigrams and unigrams is submitted in part c
As required different set of features **– [unigrams],[bigrams],[unigrams+bigrams]** (only two were asked but this report contains three) were used.
In conclusion bigrams give a slight increase in Test accuracy ~0.13 % but not that significant.