NAÏVE BAYESIAN CLASSIFIER

NAME: ASHUTOSH JAYENDRA UPADHYE

ID : 1001581542

For multi class classification, we have used use multinomial Bayesian classifier. We have data from 20 news groups which in-turn have 1000 documents each. The data of 500 documents of each class was taken as a training a set. We vectorize the words read from each file and copy it into a dictionary associated with each label.

We can also use NLTK for removing the stop words. The identifiers like “ ’’ ” and “,” can also be removed from the lists to ensure data cleaning is done properly.

The log probability of each word occurring for a particular class is calculated by getting the occurrence of a word in a class divided by total occurrences in all classes. To avoid underflow of division, a small threshold value of 0.0001 is added to the count if the word doesn’t occur in the dictionary.

Once the probability is calculated, a score is calculated for words occurring in each class and the maximised score is calculated.

The remaining 500 documents of each of the folders is used as a test data set.

The accuracy of the model is calculated by calculating the predictions which are positive per label class divided by total class labels.

Accuracy of 87 % is achieved roughly if k=1 is only used for k-fold cross validation