Assignment 5

Naïve Bayes Classifier:

In this assignment, we are given three domains for scientific articles each consisting of 300 articles. This code in performing text classification on all the three domains and checking for the accuracy. initially a set of words were collected from all the articles and store as training data. Then all the unwanted words listed in the stoplist.txt were removed from the bag of words. Arranging the bag of words in alphabetical order, a feature vector is created which takes the value 0 or 1. If the vector is 1 then the word is present in the vocabulary. Here a feature vector of size m+1 is created as to correspond to the class label of that word. Here the class labels are: ‘a’ for arxiv, ‘j’ for jdm and ‘p’ for plos, as per the category.

For the classification phase, the library used is ‘from textblob.classifiers import NaiveBayesClassifier’. Here in the testing phase, the articles are converted to testing data consisting of the feature vector (0 or 1) The difference between the training data and testing data is that testing data doesn’t have class labels. Further the training data and testing data are compared to find the accuracy of the class label for each. Hence proving the accuracy of Naïve Bayes classifier. The accuracy is shown as either 0 or 1. In this case for arxiv, the accuracy is 1 and for the rest it’s 0 as there are many words which belong to other class labels as well.