**NaiveBayes Implementation**

**CSE 6363**

**HW1**

Tasneem Burmawala

NaïveBayes Algorithm:

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), naive Bayes classifiers are a family of simple [probabilistic classifiers](https://en.wikipedia.org/wiki/Probabilistic_classifier) based on applying [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with strong (naive) independent assumptions between the features.

Language Used: Python3

Implementation of the Project:

1. Dividing into Training and Test dataset
2. Extracting bag of words in training dataset
3. Calculation on Training Data Set:

Calculation probability of every word in a class(P(W|C))

Formulae: Probability of each word in a class= count of a word in a class+1/total number of words in a class+unique words in entire dataset

1. Test Data Set:

Removed punctuations,stopwords,etc from all the files in the Test Data Set

Now, every file from the test word is taken, each word is checked in all the 20 training classes,if present it does c+=math.log(P(W|C)) of that word else c+=math.log(1/( total number of words in a class+unique words in entire dataset))

1. We now get probability of a file being in 20 classes.The class for which there is maximum value of probability is the predicted class
2. Counter is set to increment by 1 for all the match and we get the predicted class for each file
3. Now the count found in point 6 is divided by total no. of files in the test folder and then it is multiplied by 100.This is the accuracy of the correct class prediction of the testing files to the correct newsgroup class.

Result:I am getting 80.68% accuracy for the class prediction of the testing files to the correct newsgroup class.

Performance:

For my 8gb RAM , it takes about 10-15min to execute the entire code.