Naïve Bayes Classifier

**INTRODUCTION**:

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

**DESIGN & IMPLEMENTATION**:

Initially, I have divided the data into two parts, training and testing. Each folder has 450 text articles. All the data from the articles is imported into a python list and with pre-processing steps, I have removed the special characters and stop words. Now I have all the words in the training data in the list ‘bag’. After pre-processing, I have created a feature vector for each of the articles which has ‘1’ if the word is present in the bag or ‘0’ if the word is not present. This list is named ‘train\_feature’.

After the features are created, I have added one more tuple to each of the sub-list in ‘train\_feature’ which has ‘A’,’J’ or ‘P’ for arxiv, jdm or plos. This completes the manipulations on training data. Now, the same features are created for the testing data and named ‘test\_feature’. Then we come to the final classification step.

In the final step, the probabilities of each tuple is measured and stored in the ‘multA’, ’multJ’ and ‘multP’ variables. Now, each of the articles in the test\_feature is classified as ‘A’ for arxiv, ‘J’ for jdm and ‘P’ for plos depending on the maximum values of the mult variables.