Project 2. Bayesian Classifier for Newsgroup Data

Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents.

Data Set: http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html

Dataset contained 20,000 newsgroup messages drawn from the 20 newsgroups. Each of 20 newsgroups contains 1000 documents.

The data was divided into two equal halves – training data and testing data. First half of each class folder was taken as training data and rest as testing data.

The data was cleaned by removing stop words and punctuation symbols and special characters.

Naïve Bayesian was then performed on data and the accuracy was calculated on the basis of the predicted class and actual class.

The accuracy turned out to be: 83.52%

Darshil Parikh
UTA ID: 1001557968
Course: CSE 6363 – Machine Learning