# PROJECT REPORT

JAIDEEP SINGH KAINTH

## PROBLEM

Task of this project is to use Naïve Bayes Classifier to classify news text articles.

## DATA

Data is taken from the following link

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

Data is 20 Newsgroups with each group containing 1000 Documents.

## TOOLS

Language used: Python 3.5

Libraries used: os, re, nltk.corpus and math

## METHOD

We have data of 20 News Groups each containing 1000 documents. I used half of the data for training and the rest half for testing. The first step was to convert the text of the documents into numerical feature vectors by using bag of words model. I created a dictionary of unique words to store all the unique words from all the documents with the number of times that word has occurred in each group. Each document was read and was segmented into words and then these words were stored in the dictionary. I removed Stop Words like "on", "this" etc to improve the accuracy of the result.

Second step was to test the rest of the data using Naïve Bayes classifier. We try to predict the group to which the article belongs and then compare it with the given result to check the accuracy of the classifier. In this we calculate the probability of each category for that document and then categorizing based on the maximum probability. The probability is calculated by calculating the prior probability of the category which is number of articles in a given category divided by total number of articles in all the groups. Then this prior probability is multiplied by the probability of each word in a given category which is number of times that word has occurred in the given category plus one divided by total number of words in that category plus total number of unique words. But with this I got the problem of zero probability of some words due to underflow error, which makes the entire result zero. To solve this,
I used log which converts products of probabilities to sum of probabilities as follows:
$\log(P(D/y)) = \log(P(y)) + \Sigma \log(P(x_i/y))$, where $x_i$ is a unique word of Document D and Y is a particular newsgroup.

## RESULT

After Removing Stop Words, Special Characters and digits from the articles, the Accuracy I got to classify articles is 84.89697939587919%.