News Text Classifier

Project Report

**Name:** Pranav Kalambe **Email:** [pranav.kalambe@somaiya.edu](mailto:pranav.kalambe@somaiya.edu)

**Roll no.** 1913023

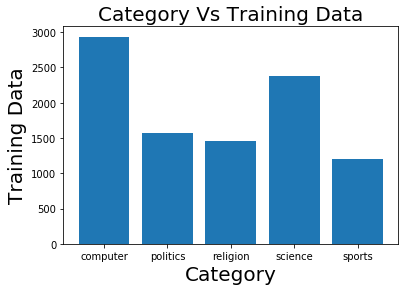
**INTRODUCTION**: The main objective of this project is to build and design a model which can take a news article as input from user and place it under different categories in news which can be further used to classify such articles in newspapers or some news websites.

**OBJECTIVE:** The given text of a news should be classified into different categories which includes sports, science, religion, politics and computer by using naive bayes multinomialNB and inverse document frequency.

**DATA ANAYLSIS:** The data set for the following problem statement has been taken from a website (data set has been submitted). The data is in text document format which has been classified into different folders named computer, politics, religion, science, sports. There are total 9537 samples.



The graph below shows the no. of text document under different categories:

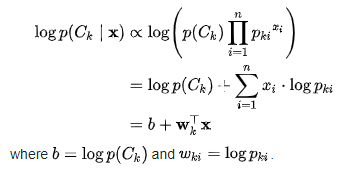


**Algorithm:** The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work. **Naive Bayes classifiers** are a family of simple “Probability Classifier” based on applying “Bayes’ Theorem” with strong (naïve) independence assumptions between the features.



Where pi is the probability that event *i* occurs

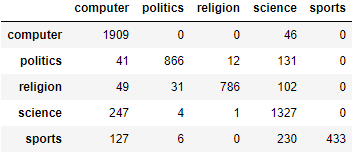
The multinomial naïve Bayes classifier becomes a linear classifier when expressed in log-space:



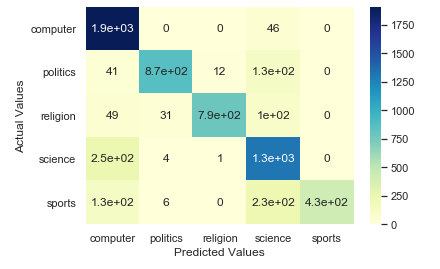
Note: If a given class and feature value never occur together in the training data, then the frequency-based probability estimate will be zero, because the probability estimate is directly proportional to the number of occurrences of a feature's value.

**Data Visualization after Training of model:**

Dataframe of Confusion Matrix



Heatmap of Confusion Matrix



**Conclusion:** The accuracy of this model based on naïve bayes multinomial NB is 83.84%. This can improved by using more techniques. The average time required for the execution is 28.363 sec (Without considering custom user input and its processing).

(Note: Dataset and Jupyter notebook has been attached with this word docx.)