**Assignment: Text Classification**

**Sifat Ahmed**

**A-1819**

Introduction

In this assignment, I have tried to analyze and classify some data collected from stack exchange. In this assignment Naïve Bayes and K Nearest Neighbours algorithm were implemented using Python, NumPy, Nltk and Pandas. After the analysis, I’ve tried to make a comparison to distinguish which classifier performed better.

Dataset

The dataset contains 12 classes with around 700 observations for each class for training. To decrease the computational complexity, I took 4 classes with 400 observations for each class. And converted to CSV from XML.

Preprocessing

The dataset for both training and testing were hardy preprocessed. For the preprocessing technique, Natural Language Processing Toolkit (NLTK) was used. The data were preprocessed in different steps:

* Tokenize to words
* Remove stop words, punctuations and numbers
* Stem the word to get the root word
* The word must contain at least 3 alphabets

KNN

For the KNN classifier, there are two parameters. One is n\_neighbours, and distance. n\_neighbours can have real positive values eg 1,2,3 . And for distance ‘Hamming’ and ‘Euclidian’ are possible values as ‘Cosine Similarity’ hasn’t been implemented yet.

Naïve Bayes

For the NB classifier, there is only one parameter which is smoothing factor. The naïve bayes represents the multinomial variation for texts. And smoothing factor can be any real number. Default smoothing factor is 1.

Comparison