**Assignment 01: Analyzing Spam Collection Data**

**Description**

**Problem:**

Analyze the given Spam Collection dataset to:

1. View information on the spam data,

2. View the length of messages,

3. Define a function to eliminate stopwords,

4. Apply Bag of Words,

5. Apply tf-idf transformer, and

6. Detect Spam with Naïve Bayes model.

**Assessment:**

(Note: For code refer the attached Notebook in pdf format)

**1. View information on the spam data**

* Import the required libraries such as pandas and string.
* From nltk.corpus import stopwords.
* Using ‘pd.read\_csv()’, get the spam data collection form ‘SMSSpamCollection.csv’ file
* Save the data in ‘df\_spam\_collection’ variable.
* Using head (), view the first five records in the dataset.
* Describe the data using describe () function.
* Since, feature name is not available, Read the file once again to include Feature name (response and message).
* Using head (), verify the dataset once again.
* View response using group by and describe method.

**2. View the length of messages**

* Verify the length of the messages and also add it as a new column in to the dataset.
* To verify it, use head () to view the records.

**3. Define a function to eliminate stopwords**

* Define a function to get rid of stopwords present in the messages.
* Check whether punctuations are present.
* Remove the punctuations from stopwords

**4. Apply Bag of Words**

* Start text processing with vectorizer, import CountVectorizer from sklearn feature extraction library.
* Use bag of words by applying the function and fit the data into it.
* Length of bag of words stored in the vocabulary\_attribute is ‘11425’.
* Store bag\_of\_words for messages using transform method.

**5. Apply tf-idf transformer**

* Apply tfidf transformer and fit the bag of words into it (transformed version).
* The shape of the tfidf is (5572, 11425).

**6. Detect Spam with Naïve Bayes model**

* Choose Naïve-Bayes model to detect the spam by Importing the MultinomialNB from sklearn.naive\_bayes.
* Fit the tfidf data into model.
* Check model for the predicted and expected value say for message#2 and message#4.