School of Computer Science Engineering and Technology

Course- B. Tech	Type- Elective
Course Code- CSET346	Course Name: Natural language processing
Year- 2023	Semester- Even
Date: 13-02-2023	Batch- ALL

Lab Assignment 02 – SMS Spam Detection

Learning Object

The main objective of this assignment is to learn the Bag-of-Words (BoWs) vectorization through a real-life problem solving. In this motive you will solve the problem of ***spam SMS detection***. In the process of doing this task, text data pre-processing, text vectorization (Bag-of-words) concepts will be implemented. Finally, the classification will be implemented with Naïve Bayes classifier.

Goal

If you have a cell phone, you probably use it dozens of times a day to text people you know. But have you ever gotten a text message from an unknown sender? It could be a scammer trying to steal your personal and financial information. Suppose you are a security analyst and you have been given a task to detect the spam SMS received by the company.

Input Dataset

Download the SMS spam collection dataset from the following link:
https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

Output

Classify a SMS is spam or not

Tasks

1. Take *SMSSpamCollection* file as input.
2. Implement the necessary pre-processing
    a. Tokenization
    b. Remove unnecessary elements using regular expression
    c. Stop-word removal
    d. Stemming
3. Create the document matrix (using Bag-of-Words)
4. Split the dataset into training and test sets.
5. Use Naïve Bayes classifier for the training the classification model.
6. Find the performance accuracy of the classifier for test data set. (Hint : Use confusion matrix for calculating the accuracy value).

Useful links:

1. https://www.datacamp.com/tutorial/stemming-lemmatization-python
2. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
3. https://scikit-learn.org/stable/modules/naive_bayes.html

School of Computer Science Engineering and Technology