THE UNIVERSITY OF LAHORE

Assignment-2
Natural Language Processing

Instructor: Muhammad Umair
Points: 10
Deadline: 27-09-2023

Instructions:

- Late submissions would not be entertained.
- Plagiarized work would not be entertained.
- Use Python as a programming language to demonstrate the assignment work.
- Apply preprocessing on at least two datasets.
- Submit PDF print of Jupyter Notebook Named as Student_Name_Reg#
- Mention your Colab Notebook shareable link on the top of your submitted PDF.

Q1: Text data is available to a great extent which is used to analyze and solve business problems. But before using the data for analysis or prediction, processing the data is important.
To prepare the text data for the model building we perform text preprocessing. It is the very first step of NLP projects. Some of the preprocessing steps are:

1. Removing punctuations like., ! $( ) * % @
2. Removing URLs
3. Removing Stop words
4. Lower casing
5. Tokenization
6. Stemming
7. Lemmatization

Apply these preprocessing steps on given datasets or you can take any public datasets of your choice.

Datasets:

- Large Movie Review Dataset
- Sentimental Analysis for Tweets
- E-Mail Spam classification

Reference Material:

- Natural Language Toolkit
- Google Colaboratory as Python IDE
- Regular Expression Documentation
- W3School for Sample RE Implementation
- Build and Test RE Online