Lab Assignment – 1 (Part 3)

Objective: Text data pre-processing.

1: **Download** at least 60 documents belong to different formats such as txt, doc, pdf, html etc and keep it in a single folder.

2: **Read** the text documents using different parsing mechanisms and keep them in an array.

3: **Standardize** the dataset using following pre-processing techniques.

- Remove all the special characters, smileys, and keep only alpha numeric values in the txt.
- Convert upper case letters into lower case letters.
- Remove the words which are not in this English dictionary with half million words https://raw.githubusercontent.com/dwyl/english-words/master/words.txt
- Store the pre-processed data in another file (You may store it in .txt or .csv or xlsx)

Suggested Platform: Python: Azure Notebook/Google Colab Notebook, packages such as numpy, nltk, regular expression package re.

Marking: Marking is based on both **performance during the lab hours** as well as **complete submission**.