CS5601 Natural Language Processing
Dr. Koninika Pal
Gayathri R, Mohammad Taufeeq
IIT Palakkad
Assignment 1: Hand out 1.09.2023, Due 11.09.2022

INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD

Assignment 1: Regular Expression

Extract the dates mentioned in the given input text file, **NLP_Exercise1_Input.txt**, and standardize to a unique format. Once a date is identified it should be standardized to **day-month-year** format (note the hyphen). For example, if your program identifies **5/08/2023** as a date, it should standardize it to **5-August-2023**. If the identified date does not have an explicit year then the default year should be 2023. For example, **17 August** in the input text should be extracted as **17-August-2023**

Assignment 2: Sentence Segmentation

Implement sentence segmentation on the same input file **(NLP_Exercise1_Input.txt)** using different methods.

(a) By using rule-based methods (For example, take punctuation marks as delimiters)

(b) By using libraries such as SpaCy and NLTK.

In either method, explain your approach in brief, use appropriate evaluation metrics, and report your results.

Assignment 3: Heap's Law

Download a book of your choice from the Gutenberg project. For example, you can find The Adventure of Tom Sawyer in this link.

(a) Calculate the running TTR for every 2000 words and note down your observation.

(b) Plot the total vocabulary changes with the advancement of the chapters(you can plot it using the total number of tokens).

(c) Segment the sentences and find the POS tag for words using Spacy or NLTK library. Plot how the vocabulary changes for the following three types of words with the advancement of the chapters.

  (i) Noun (consider all NN tags)

  (ii) Verb (consider all VB tags)

  (iii) Every POS tag except nouns and verbs