MAI372– Natural Language Processing

III MSc AIM

16-02-2024

Regular lab Question

Lab Exercise 2:

1. Write a paragraph based on your interested Domain and that should incorporate the special characters, punctuations, stop words, negation (don't), Email Id, Phone Number and Date of birth (27-March -2000). Perform the following operations:
    a. Count the number of unique tokens in the text.
    b. Count the number of punctuation characters and remove punctuations.
    c. Plot the distribution of stopwords used in the text using barplot.
    d. Remove the stopwords from the text.
    e. Plot the distribution of each POS Tag using a barplot.
    f. Determine the number of unique lemma available in the text.
    g. Plot the frequency distribution of words in the text. Display only the top 10 (most occurring) tokens in the chart.
    h. Find the number of unique bigrams, trigrams and quadgrams (n=4) in the corpus
    i. Find all dates and convert them to the DD-MM-YYYY format.
    j. Plot a distribution of the different values of year occurring in the text.
    k. Determine whether the text contains any phone numbers in it. Ensure the phone numbers are valid if any. If an invalid phone number is found, remove it from the text.
2. Write note on insights and the possible applications (According to your Knowledge) for the aforementioned applications.

Program Evaluation Rubrics

C1-Timely Submission---2 marks,

C2-Correctness & Clarity---4 marks,

C3-Complexity & Validation---2 marks,

C4-Viva Voice------2 marks

General Instructions

1. The file you have to save with your name, last 3 digits of register number and program number "Aaron_201_Lab1".
2. The implemented code you have to download and upload in the Google Class room in the given scheduled time.