MAI372– Natural Language Processing

III MSc AIM

10-02-2024

Regular lab Question – I

Lab Exercise 1:

1. Write a paragraph based on your interested Domain and that should incorporate the special characters, punctuations, stop words, negation (don't) and emojis (try to add more than one emoji 😊🥰😍 contiguously in sentence). Perform the following types of Tokenization and utilize the Python libraries to tokenize.
    a. Word Tokenization
    b. Sentence Tokenization
    c. Punctuation-based Tokenizer
    d. Treebank Word tokenizer
    e. Tweet Tokenizer
    f. Multi-Word Expression Tokenizer
    g. TextBlob Word Tokenize
    h. spaCy Tokenizer
    i. Gensim word tokenizer
    j. Tokenization with Keras

NB: Write note on insights and the possible applications (According to your Knowledge) for the aforementioned Tokenizers.

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.