Lab3

Basic Text Processing and Regular expressions

Date: 09-03-2021

Objective

The objective this lab is to familiarize with basic text processing and identifying patterns from text using regular expressions.

Instructions

- Use Jupyter/Colab notebook for implementation
- Write your observations in the notebook as markdown text
- Convert the notebook to PDF
- Submit the PDF before due time.

Steps

1. Access a webpage using Python (use package urllib)
2. Remove HTML tags using regular expression
3. Identify the following patterns:
    - Email id: id@domain
    - IP address
    - Phone number
    - Website address
    - Abbreviations (all capitals)
    - Proper nouns (name of company/person/places)
    - Numbers
    - Dates
4. Prepare the set of unique words
5. Now create a new vocabulary after:
    - Removing stopwords
    - Treating words irrespective of cases (upper/lower)

    - Applying stemming
6. Compare the set of words in steps 4 and 5. How much reduction have you achieved?

References:

1. [NLP] Basics: Understanding Regular Expressions | by Céline Van den Rul
2. Chapter-3: NLTK Book (NLP with Python): https://www.nltk.org/book/ch03.html
3. Chapter-2 (Regular Expressions), SLP by D. Jurafsky: https://web.stanford.edu/~jurafsky/slp3/2.pdf
4. Python | Stemming words with NLTK