CSE 3024: Web Mining **Slot: L51 + L52**

Online Submission Deadline: 08th August 2020

Fundamentals of NLP and Crawling

[3 + 3 + 2 + 2]

- **Upload your code and result as a single PDF file in VTOP and MOODLE**
- **File should contain**
  - **Question**
  - **Code**
  - **Result / Output screen**

---

1. Write a python program to
   a. Extract the source content (excluding any tags) from the website (https://en.wikipedia.org/wiki/Web_mining).
   b. Display the total number of terms and term frequency of each term present in them after applying stop word removal.
   c. Remove all the special characters/symbols present in the content by adding those characters as stopwords in the existing stopword list..
   d. Also, apply stemming (don't use porter stemmer) and lemmatization to the same document and display the number of terms along with their corresponding stemmed as well as lemmatized words present in them using Pandas dataframe as per the format given below:.

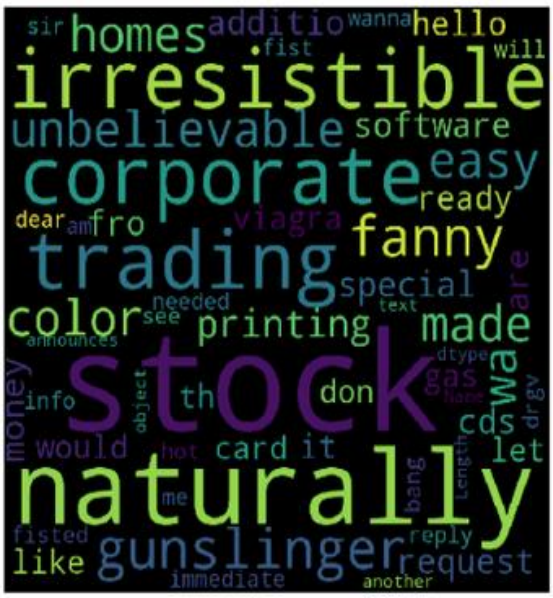
Original Term	Stemmed Term	Lemmatized Term
Studied	Stud	Study

   e. Count the total number of stemmed and lemmatized words.
   f. Display the POS tag (sentence-wise) for all the stopwords (excluding the special character/symbols), which are removed from the content, using pandas dataframe as per the format given below:

Original Sentence	List of Stopwords	POS-Tags
Web mining is the application of data mining techniques to discover patterns from the World Wide Web.	is the of to from the	VBZ DT IN TO IN DT

2. Write a python program to
   a. Extract the contents (excluding any tags) from two websites (https://en.wikipedia.org/wiki/Web_mining&https://en.wikipedia.org/wiki/Data_mining).
   b. Remove stopwords (including the special characters/symbols) from the contents retrieved from those two URLs and save the contents in two separate .doc file.
   c. Display the Term-Document incidence matrix using Boolean, Bag-of-words and Complete representation (Use pandas dataframe).
   d. Input a search a query (preferably a sentence) and compare the contents of the both pages with the processed query. Display the similarity result based on highest frequency matching count of the term.

3. Write a python program to prepare the **Word Clouds** representation based on the content present in the two document files prepared in Q.No. 2. A sample Word Clouds representation is provided below for reference.



4. Write a python program to show the implementation of sentence paraphrasing through synonyms (retaining semantic meaning) for the following four sentences. Display at least three other paraphrased sentences for each sentence mentioned below.

   a. The quick brown fox jumps over the lazy dog
   b. Obama and Putin met the previous week
   c. At least 12 people were killed in the battle last week
   d. I will go home and come back tomorrow.