CSE 3024: Web Mining **Slot: L39 + L40**

Online Submission Deadline: 23rd February 2021

Fundamentals of NLP and Crawling

[4 + 3 + 3]

- **Upload your code and result as a single PDF file in VTOP [Mandatory] and MS Team Assignment [optional] on or before the deadline.**
- **No other form of submission will be acceptable.**
- **If you failed to upload in VTOP on or before the deadline, but successfully uploaded in MS Team Assignment, then 2 marks of penalty will be imposed on the secured marks.**
- **If you fail to upload your assignment in both VTOP and MS Team Assignment, then your assignment will not be evaluated and ZERO (0) mark will be awarded.**
- **File should contain**
  - **Question**
  - **Code**
  - **Result / Output screen**

---

1. 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 [using Spacy Module] (including the special characters/symbols) from the contents retrieved from those two URLs and save the contents in two separate .txt file.
      - [List of additional Stop words to be considered = [dot, comma, single-quote, double quote, question mark, brackets [square, parentheses, curly, angle], exclamation mark]]

   c. 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:

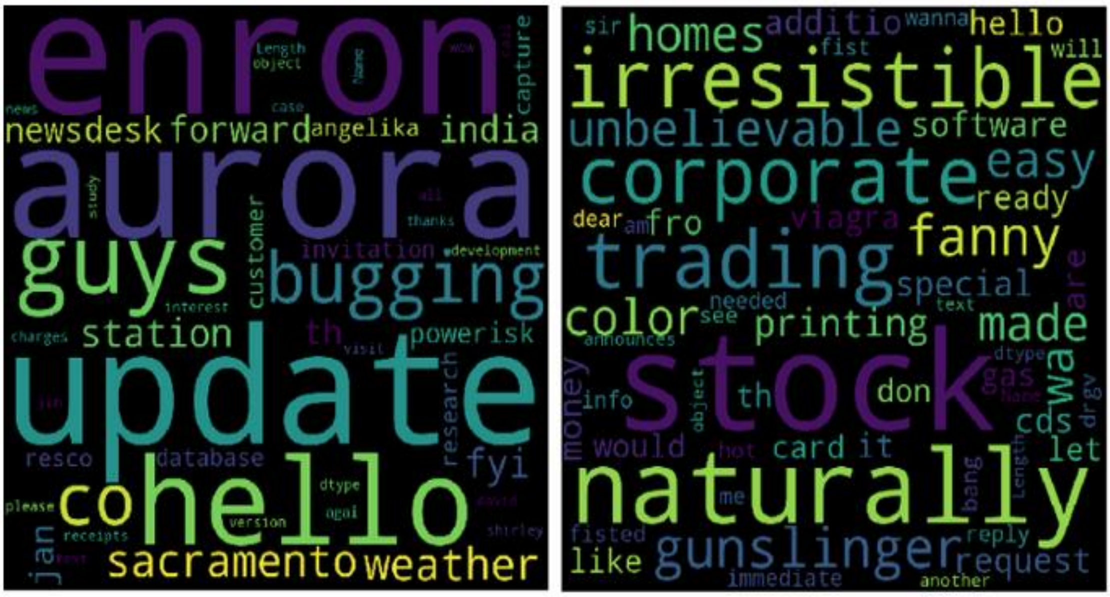
Original Sentence	List of Stopwords	POS-Tags
Web mining is the application of data mining techniques to discover patterns from the World Wide Web.	is	VBZ
	the	DT
	of	IN
	to	TO
	from	IN
	the	DT

d. Display the Term-Document incidence matrix using Boolean, Bag-of-words and Complete representation (Use pandas dataframe). Prepare three separate table, one for each type of representation as per the format given below:

Terms	DOC1	DOC2
Web	5	0
Data	0	1

e. 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.

2. Write a python program to prepare the **Word Clouds** representation based on the content present in the two document files prepared in Q.No. 1. A sample Word Clouds representation is provided below for reference.



3. 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. We can rewrite history as much as we like.
c. Once you know all the elements, it’s not difficult to pull together a sentence.
d. The incessant ticking and chiming echoed off the weathered walls of the clock repair shop.