NLP: Assignment 1: (Prof. Ashwini M. Joshi)

Date: 05-02-2020 **Date of Submission: 10-02-2020**

Part 1

Use NLTK library of Python. NLTK library has several corpora (text datasets) one of which is the "Inaugural Speeches of all the US Presidents" from 1789 - present

A. Extract the raw text from the latest 5 speeches

B. Tokenize the raw text of these speeches into sentences

C. Tokenize the each sentence into words

D. From the tokenized sentences, generate new sentence which is the sentence formed by removing the stop--words as well as stemming each word in the sentence.

E. Tokenize the new sentence in D into words.

F. Make separate Word Clouds of the words in C and words in E.

G. What inference can you draw from the two word clouds? Does the stemming and stop--word removal give you any different insight into the corpus?

H. Using POS tagging identify the frequency distribution of the different parts of speech of the words given in the text. (Use PennTree Tagset). Identify and represent the distribution via suitable visualization technique.

I. Each sentence can be scored as the sum of frequencies of individual words (excluding stop words). Identify the top 5 sentences based on this score.

J. With the help of regular expressions, extract the words in E that have either numbers or other non--alphabetic characters.

Part 2

Go through the TextBlob library (an API over NLTK) @ https://textblob.readthedocs.io/en/dev/.

TextBlob has two taggers i.e. Pattern Library and NLTK Tree Bank tagset. Use TextBlob for NLTK tagger. The default tagger of TextBlob uses Pattern implementation
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Use TextBlob to Tag the word tokens and choose the tokens according to tags. You should install Python word cloud library (conda install). Note that it requires a string of words.
https://github.com/amueller/word_cloud

In case of windows, you may have issues and will like to go ahead with the unofficial windows binary from here
https://www.lfd.uci.edu/~gohlke/pythonlibs/#wordcloud

Make word clouds with the followings. Try to make the most well-meaning word clouds showing the noun word cloud vs. adjective word cloud

- Nouns

- Adjective

Note: Submit the Python Notebook to amj.nlp2019@gmail.com clearly mentioning your team number and team member details (SRN and Name) at the top of the notebook