Homework Assignment-1

Due Oct 11 by 11:59pm **Points** 50 **Submitting** a text entry box or a file upload
Available Oct 5 at 7pm - Oct 11 at 11:59pm 6 days

This assignment was locked Oct 11 at 11:59pm.

Part-1 - 35

- (3) Import a book from the Gutenberg Project in NLTK, and tokenize the text.
- (5) Compute the vocabulary of the book. To do that, you will need to find the frequency distribution of tokens. Save the distribution in a CSV file using the format: token: frequency
- (5) Next, determine the POS tags for the book's entire text, and find the frequency distribution of the POS tags.
- (4) Plot the *cumulative* frequency distribution of the most frequent tokens, and the *simple* frequency distribution of the POS tags.
- (3) Investigate the difference between the tokenized version of the book (already provided by NLTK) and the tokenization using NLTK's word_tokenize function. Write your finding in a text file
- (8) Use a corpus for Names or Open source tool (e.g., Spacy) to find the person names in the book and output the most frequent name.

(7) For doing the above steps, you will need to follow these:

(2) Sort tokens (and POS tags) by frequency.

(1) Limit the number of the most frequent tokens to 50.

(2) Remove tokens containing non-alphabetic characters.

(2) Remove stop words.

Briefly summarize your findings of the tokenization differences in a separate text file.

Part-2 - 8

(1) Import Reuters corpus (all the documents in the corpus) from NLTK.

(1) Find frequency distribution of the words.

(2) Plot the frequency distribution.

(1) Find top-10 words according to frequency.

(3) Now find the frequency distribution of the topics. Plot the frequency distribution (limit to 10 topics). Find top-10 topics according to frequency.

Part-3 - 7

(2) Import Twitter Corpus from NLTK.

(3 - 1 pt for each) Find the total number of Hashtags, Mentions and URLs in the corpus.

(2) Now remove the Hashtags and Mentions from the tweets and output the cleaned tweets

**** Name your submission file something like this ****

nlp220-assignment1-cruzid.py

Also name your zip file as per this format