NLP Homework 1

Due Monday, September 21, 11:59 pm.

Corpus Statistics and Python Programming

For this assignment, please read Chapter 1 and 2 of NLTK book carefully.

With the COVID-19 pandemic, people talk about issues related to Coronavirus in various online platforms. These texts help us better understand people's views and their concerns, and things they care about. As a starting point, you will analyze a subset of text from news/message boards/blogs about CoronaVirus.

1. **Data**

The data you will analyze in this assignment includes two .json files from the free dataset from news/message boards/blogs about CoronaVirus. This dataset has four months data – 5.2 M posts. The time frame of the data is Dec/2019 - March/2020.

Please check out the web site for more information about this dataset. Please describe the characteristics of this corpus briefly, e.g., who created this dataset, the format of .json file and the fields, the related policy (if any), etc. (no more than 200 words).

The two .json files that you will analyze are:
16119_webhose_2020_01_db21c91a1ab47385bb13773ed8238c31_0000002.json
16119_webhose_2020_02_db21c91a1ab47385bb13773ed8238c31_0000002.json

2. **Data Pre-processing (20%)**

You will code in Python to extract the content of the following fields and save it to a CSV file: "facebook":{…}, "title", "published", "replies_count", "author", "url", "country", "text".

Note: regarding the "text" field, you will decide how to process the words, i.e. decide on tokenization and whether to use all lower case, use or modify the stop word list, or lemmatization. Briefly state why you chose the processing options that you did.

3. **Data Analysis**

You will first get a summary of the data, e.g., the number of entries, the average length of the text, the countries, etc. You are encouraged to use graphs to present your summary.

To get a rough understanding of what these texts were about, you will also perform the following three tasks on the pre-processed data:

- list the top 50 words by frequency
- list the top 50 bigrams by frequencies, and
- list the top 50 bigrams by their Mutual Information scores (using min frequency 5)

4. Interpretation of the Results

Please explain what you have learned about this data, based on the results above. And, please discuss what additional analysis tasks that you think are important to conduct and why.

How to Submit Homework:

Go to the Blackboard system and the Assignment for Homework 1 and submit your report. Your report should include:

1) Description of data pre-processing (with Python processing screenshots in the corresponding section)
2) The results from the analysis tasks (with Python processing screenshots in the corresponding section)
3) Your interpretation of the results and the additional analysis you suggest to perform in the future

Please also upload your Python code besides the report