NLP Homework 1

Due Sunday, February 18, 11:59 pm.

Corpus Statistics and Python Programming

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

You will analyze a collection of the State of the Union Addresses contents. The State of the Union Addresses dataset is a collection of annual speeches delivered by the presidents of the United States, from George Washington to Barack Obama, to a joint session of the United States Congress for the span of 1790-2016. This dataset contains the two texts combined, which are small subsets of the Project Gutenberg Ebook corpus. Here is the Project Gutenberg Ebook web page: http:// http://www.gutenberg.org/

For the following tasks, you will use a modified version of the State of the Union Addresses dataset. It contains three files, which are 1) the addresses delivered between 1790 and 1860, 2) the address delivered between 1946 and 2016, and 3) a policy description of the Project Gutenberg Ebook ('state_union_policy').

1. **Analysis of State of the Union Addresses dataset: Description (10%)**

First, review the dataset and describe the characteristics of this corpus briefly such as its history, the naming convention of its files, number of documents it contains, the related policy, etc. (no more than 450 words)

2. **Analysis of State of the Union Addresses dataset: Part1 (30%)**

In this problem, you will analyze the "state_union_part1.txt" that contains presidential addresses, based on what is covered in Chapter 1 and 2 of the NLTK book. The document is available for download in the Assignment folder in the course web site. To do so, you will:

A) Perform the following three tasks (30%, 10% for each task):
   - list the top 50 words by frequency (normalized by the length of the document)
   - list the top 50 bigrams by frequencies, and
   - list the top 50 bigrams by their Mutual Information scores (using min frequency 5)

Note: 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. Analysis of State of the Union Addresses dataset: Part2 (30%)

In this problem, you will analyze the "state_union_part2.txt" that is available for download in the Assignment folder in the course web site.

A) The analysis tasks are as follows (30%, 10% for each task):
   - list the top 50 words by frequency (normalized by the length of the document)
   - list the top 50 bigrams by frequencies, and
   - list the top 50 bigrams by their Mutual Information scores (using min frequency 5)

Note: 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.

4. Comparison (30%)

Please compare the analysis results from question 2 and question 3.

A) How are state_union_part1 and state_union_part2 similar or different in the use of the language, based on your results? Why?
B) Are there any problems with the word or bigram lists that you found? Could you get a better list of bigrams?
C) How are the top 50 bigrams by frequency different from the top 50 bigrams scored by Mutual Information?

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 results in PDF format.
2) Output (included in an appendix)
3) Python processing screenshot (included in an appendix)
4) Your Python code (submit in one separate folder zipped)