NLP Homework 1

Due Sunday, February 17, 11:59 pm

Corpus Statistics and Python Programming

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

1. **Analysis of Wikipedia discussion forum (30%)**
   In this problem, you will analyze a small subset of Wikipedia's Article for Deletion (AfD) discussion content. In a Wikipedia AfD discussion, users offer their opinions on how to handle the Wikipedia article being discussed – *to keep it in Wikipedia, to delete it from the site, to merge it with another article, etc.* Here is what a Wikipedia AfD discussion page looks like: https://en.wikipedia.org/wiki/Wikipedia:Articles_for_deletion/Log/2016_November_25#K1_Speed. This page contains 91 AfD discussions.

   In this problem, you will analyze the "HW_WikipediaDiscussions.txt" that contains the users' comments from about 40, 000AfD discussions, 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. 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.

2. **Analysis of NPS Chat corpus (25%)**
   In this problem, you will analyze the NPS Chat corpus (nltk.corpus.nps_chat) from the NLTK package. To do so, you will:
   - first review and describe the characteristics of this corpus briefly such as its history, the naming convention of its files, number of documents it contains, etc. (10%, a short paragraph, no more than 350 words)
     You can refer to online resources such as http://faculty.nps.edu/cmartell/NPSChat.htm and http://www.nltk.org/book/ch02.html for this task. You can paraphrase some of the descriptions from these resources, but plagiarism is prohibited.
   - next, you will (15%, 5% 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. **Comparison (30%)**

Please compare the analysis results from question 1 and question 2.

a) How are Wikipedia discussions and NPS chats similar or different in the use of the language, based on your results?

b) How are the processing options similar or different for the two analysis tasks?

c) Are there any problems with the word or bigram lists that you found? Could you get a better list of bigrams? How are the top 50 bigrams by frequency different from the top 50 bigrams scored by Mutual Information?

4. **Word and Name Puzzle (15%)**

hint: NLTK book examples are available: 2. Accessing Text Corpora and Lexical Resources (section 4)

E	G	B	D
A	F	K	J
L	M	O	R
C	N	S	T

How many **words** of six letters or more can you make from these letters shown above? Each letter may be used once per word. Each word must contain the letter in the green cell. Please use Python to solve this problem. Please make sure each word is in the vocabulary of nltk.corpus.words.

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)