NLP Homework 1
Due Thursday, September 26, 2013 by midnight

Comparing Corpora with Corpus Statistics.

For this homework, select or make two documents. You can use books from the Gutenberg project already provided by NLTK, the corpora in the nltk.book package, or you can choose large documents of your own. (Chapter 3 of the NLTK book explains how to read raw text from a file, among other ways to get a corpus into Python, and we will work on this in lab.) Try to pick two documents that are different in character in some aspect: generally either topic, style, genre or some cultural aspect.

You may choose to work in groups of 2-3 people, or you may work on your own. In the following list of tasks, there are 2 that are required for everyone – running the corpus statistics with a brief discussion of the ones you chose and stating a question on comparing the difference between the documents. For every person in your group, you must select an additional task, i.e. one person chooses one additional task, two people choose two additional tasks, etc. Another option for a 2 person group is to choose an additional task and an additional document for a 3-way comparison, and a 3 person group could do something similar.

1. (Required) Examine the text in the documents that you chose and decide how to process the words, i.e. decide on tokenization and whether to use all lower case, stopwords or lemmatization. Using the programs developed in the lab,
   - 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.

Note that you may wish to modify the stop word list, based on your question in Task 2.

Give a brief analysis of this data that discusses the following questions.

a. Briefly state why you chose the processing options that you did.
b. Are there any problems with the word or bigram lists that you found? Could you get a better list of bigrams? If so, speculate on how could you do this and why?

2. (Required) Describe a problem or question that is based on the difference between the two documents. In the case of literary works, for example, this could be how to characterize the style between two authors or two works of different classes. Another example would be to compare the informal text in blogs with more formal text. Or you can do a topic related comparison that selects words (as in the SOTU speeches example). You could also make a comparison of similar text but at two different times.

3. (Additional) Collecting your own data. (Of course, you do the collection before you actually do Parts 1 and 2.) Find your own documents or collect data from other sources. Combine the text from these sources to make two documents for the corpora for the first task. Describe the method that you used to define and collect the data, including the difference between the

documents. Note any limitations to the method or the text that you were able to find. Do preprocessing to get the text in a suitable format for processing and describe what you did.

4. (Additional) Answer the question stated in part 2 by giving a discussion of the comparison of the texts. Using one or more of the types of measures that you ran in the first task, i.e. word frequencies, bigram frequencies, or bigram mutual information, make a comparison of the two documents to answer the problem or question. You may wish to hand pick out particular examples of word frequencies, bigram frequencies or mutual information scores that contribute evidence for your comparison, or combine examples into categories.

5. (Additional) Read the Mutual Information paper by Church and Hanks, and write a function for their Association Ratio to have a larger window, where the window size can be specified as a parameter. Rerun your Mutual information data from Part I with a window size of 5 and briefly discuss the results.

What to submit for Homework:

Write a homework report that tells what documents you used and describes the process that you used to process it, particularly mentioning any variations from the process described in class. Present the results from Tasks 1 and 2, and write a description of the additional task(s). If you worked in a group of two or more people, describe the role(s) of each person in carrying out the tasks.

How to Submit Homework:

Go to the Blackboard system and the Assignment for Homework 1. Each person in each group should submit the report. (Even though the reports are the same for each person in 1 group, the Blackboard gradebook works better if each student submits something.) If you did any programming or used other resources (that are not too big), attach these as documents as well.