NLP Homework 1
Due Wednesday September 30, 2015 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, you can choose large documents of your own, or you can put together groups of smaller documents to make two large documents out of the corpora.

Try to pick two documents that are different in character in some aspect: generally either topic, style, genre or some cultural aspect. The work in this assignment is to run word frequencies, bigram frequencies and mutual information scores on the two documents. Then you will select items from these lists to make a comparison between the documents to answer some question about the differences or similarities between them.

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.

Note that you can choose private documents – you will not have to hand in the original documents, but you will have to hand in the word and bigram lists.

1. Choosing the data: either
   a) Choose existing large documents from NLTK or from the Gutenberg collection on the web, or
   b) [Counts as additional task] Collect your own data, by using your own documents or collecting 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.

2. [Required task] 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 process 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 3. To complete this part:

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? How are the bigram frequency list and the bigram Mutual Information lists different?

3. [Required task] 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.

4. [Additional task] Now answer the question you have chosen 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. For this analysis, you will want to choose or to revise data that will be applicable for your 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. If your documents in task 1 required a lot of preprocessing steps, you may give a short discussion here.

5. [Additional task] 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. Run your mutual information scores with your hand-written program 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 and submit your report. You may optionally also attach supplementary documents such as the python processing that you did.

Ideas for Homework

For documents that come from NLTK corpora, read the NLTK book sections from Chapter 2 on the different corpora. Also note that Chapter 2 discusses how to load your own text with the PlainCorpusReader, or you can just read text from files, which we will do in the next lab.

Example of using word frequencies to analyze text:
Nate Silver's analysis of State of the Union Speeches from 1962 to 2010.
http://www.fivethirtyeight.com/2010/01/obamas-sotu-clintonian-in-good-way.html
Here is a statement of the question that he is trying to answer by looking at word frequencies to compare the SOTU speech in 2010 with earlier speeches:
"What did President Obama focus his attention upon and how does this compare to his predecessors?"

How is this example different from your assignment? It is different because it does more comparisons that you are required to do and only uses word frequencies while you must also look at bigram frequencies and mutual information. But if you choose the analysis option, this is the type of thing that you are aiming for. Also, your do not have to make colored graphs, but you can just use word lists and frequencies/scores.

Also note that you can collect your own text from just about anywhere. For a (too short) example of this, see the potato chip marketing example.