NLP Homework 1
Due Tuessday, September 26 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.

1. Choosing the data: either
   a) Choose existing large documents from NLTK or from the Gutenberg collection on the web, or
   b) 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. 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 (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 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 top 50 bigrams by frequency different from the top 50 bigrams scored by Mutual Information?
   c) If you modify the stop word list, or expand the methods of filtering, describe that here.
   d) You may choose to also run top trigram lists, and include them in the analysis in part 3.

3. 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.

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.

Make sure you include the following in your report:

a) Clearly describe the problem or question you are trying to address through the comparison between the two selected documents.
b) Present and explain insights or conclusions based on the comparison to answer the question (do not just report numbers).

What to submit for Homework:

Write a homework report that gives the lists that you ran, and answers the discussion questions given for each of the three 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.

If you are interested in comparing product reviews, there is an Amazon product review collection by Jure Leskovec of Stanford at https://snap.stanford.edu/data/web-Amazon.html. The reviews are collected by type, and I'm not sure how big each collection is. It will take some processing to get all the reviews of one type read and the text collected into one file; near the end of the page is a python script that parses one file to find the text for you.

If you choose documents from social media, e.g. Twitter or Facebook posts, use the NLTK twitter tokenization functions.

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, you 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. Another examples is 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. From Wikipedia, we have collected the file "comment.txt" that contains the users' comments from about 40, 000AfD discussions. This file could be used as one of your documents.