NLP Homework 1
Due Tuesday, October 4, 2011

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: topic, style, etc.

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 and doing the analysis of the difference between the documents. For every person in your group, you must select an additional task, i.e. one person chooses one task, two people choose two, etc.

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 most frequent collocations (aka bigrams) and their frequencies, and
   - list the top 50 collocations by their Mutual Information scores.

Give an analysis of this data that discusses the following questions.

a. Does the collocations list from each corpus provide a good representation or characterization of it? What is the nature of the differences among the lists produced from each group?

b. Are there any problems with the collocation lists that you found? Could you get a better list of collocations? If so, speculate on how could you do this and why?

2. (Optional) Collecting your own data. 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.

3. (Optional) Do a more detailed analysis of the corpus statistics lists.
Try experiments to run the bigram frequencies or mutual informations by changing some of the following:
Running with and without stopword lists.
Running with different stopwords.
Running with and without punctuation in the bigrams.
Running with and without lemmatization.
Or another experiment of your own devising.

Can you analyze how fast the mutual information drops in value? What are some of the collocations at the bottom of the list?

4. (Optional) 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.

5. (Optional) Choose a different measure for corpora, such as Mutual Information on collocations (noting that the pairs of words can be in either order, according to some definitions). Write a function that implements this measure and run it on the corpora that you used in Part I.

6. (Required) Analysis. 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). Using one 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.

What to submit for Homework:

Write a full homework document that tells what corpora 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 Task 1, write a description of the optional task(s) and write the analysis described for task 6. If you worked in a group of two or more people, describe the role(s) of each person in carrying out the tasks.

Also prepare a short description of your corpora and a short version of any interesting analysis or results that can be shared with your classmates.

How to Submit Homework:

Go to the Blackboard system and the Assignment for Homework 1. One person in each group should submit your full document. If you did any programming or used other resources (that are not too big), attach these as documents as well.

Go to the Homework 1 discussion list and post your short document for the rest of the class.