**SDA/LING 250 – COMPUTATIONAL TEXT ANALYSIS**

Department of Linguistics - Simon Fraser University – Dr. Maite Taboada **Assignment 1**

**NLTK and corpus functions**

This is a group assignment. You need to work in groups of 2-3 students, to be organized through Canvas. Please work together and submit only one assignment for the group.

Please list the group members’ names and student IDs on the first page, together with a list of responsibilities (sample below). Make sure that, as an individual in the group, you understand and agree with the answers submitted.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name (Last, First)** | **Student ID** | **Section contributed** | **Section edited** | **Other contributions** |
| **Jiang Long** | 200099436 |  |  |  |
| **Antanila H.** | 301332035 |  |  |  |
| **Sava Savkovic** | 301397121 |  |  |  |
|  |  |  |  |  |

Create or find your own corpus. This corpus should have the following characteristics:

* Three different subcorpora, belonging to three different genres (registers or text types)
* Each genre at least 5,000 words long

You have various options. You can collect your own email or text messages from several days, your own papers or other course work, or you can collect from external sources. A few things to take into account:

* If you use email or text messages, you can only include text sent by you. Text by other participants (including threads in quotes in your own messages) needs to be excluded for confidentiality reasons. Alternatively, you can ask the people involved to give you permission to use the data for research purposes. This also includes text posted on social networks, where a password or permission is required to access the data.
* If you use text from web sites, you do not normally need permission, unless the site is protected under a password or similar.
* Newspaper text is certainly the easiest to find, but please make sure you have distinct genres within it. For instance, you could use a high-brow vs. low-brow newspaper, or opinion pieces vs. news articles.
* You will need to save the corpora as plain text.
* It is fine to use all written text, but if you have access to transcribed sources of spoken language, feel free to use that as well.
* Please include a reference to the source of each genre in your assignment.

You can work through the Assignment 1 prep available on Canvas and on GitHub classroom to work through some of the tasks below.

1

**Here is what you need to submit, for each subcorpus**:

1. The length (in words).
2. The lexical diversity.
3. The longest sentence (type the sentence and also give the number of words). Hint: look at the Gutenberg part of Section 2.1 in NLTK.
4. The top collocations.
5. The top ten words that start with each of the vowels (involves using FreqDist).
6. A stemmed version of the longest sentence (extracted above in 3).

For the assignment, you need to submit:

* The answer to the question.
* The commands used to arrive at the answer.

You can submit one notebook with the commands and the output or submit the commands and the output in a text/word/pdf file.

# Notes Note 1

In order to perform any operations on the actual content of the corpora, you need to assign the entire content to a variable. Let’s say that you did the following when you loaded:

corpus\_root = 'C:\my\_data'

mycorpus = PlaintextCorpusReader(corpus\_root, '.\*')

That is, you named the collection of texts in the directory “corpus” mycorpus.

Now you need to assign the *words* of each text to a variable. If each of the three corpora are called, for instance, news.txt, opinion.txt, and letters\_to\_the\_editor.txt, you’ll need to do:

news = mycorpus.words('news.txt') opinion = mycorpus.words('opinion.txt') letters = mycorpus.words('letters\_to\_the\_editor.txt')

This will assign to the variable news all the words in that file. Now you can use the regular commands that you see in the book with “text1”, etc., using instead “news”. See Chapter 2 of the NLTK book.

# Note 2

Functions such as text1.concordance() work only with the texts pre-loaded with NLTK. To use them with other texts, you have to again define the text into a variable, using a function such as the following (for each corpus):

news = nltk.Text(mycorpus.words('news.txt'))

See the same section above on the Gutenberg corpus for more detail.

2