## HU Extension Assignment 12 E63 Big Data Analytics

### Handed out: 04/23/2016 Due by 11:30PM EST on Friday, 04/29/2016

Please, describe every step of your work and present all intermediate and final results in a Word document. Please, copy past text version of all essential command and snippets of results into the Word document with explanations of the purpose of those commands. We cannot retype text that is in JPG images. Please, always submit a separate copy of the original, working scripts and/or class files you used. Sometimes we need to run your code and retyping is too costly. Please include in your MS Word document only relevant portions of the console output or output files. Sometime either console output or the result file is too long and including it into the MS Word document makes that document too hard to read. PLEASE DO NOT EMBED files into your MS Word document. For issues and comments visit the class Discussion Board. If you use some other language other than Python in your daily work with NLP, please be free to use that language and a framework of your choice to do this assignment.

**Problem 1.** Create atable displaying **relative** frequencies with which “modals” (can, could, may, might, will, would and should) are used in 18 texts provided by NLTK in their extract from Gutenberg Corpus. For two modals with the largest span of relative frequencies (most used minus least used), select a text which uses it the most and the text that uses it the least. Compare usage in both texts by examining the concordances of those modals in two texts. Perhaps try to understand how are those words used in different texts.

from nltk.corpus import Gutenberg

execfile("/Users/rpulekar/problem1\_relative\_frequencies.py")

can relative frequency span: 0.24-0.06 = 0.18

could relative frequency span: 0.45-0.03 = 0.42

may relative frequency span: 0.12-0.05 = 0.07

might relative frequency span: 0.17-0.02 = 0.15

will relative frequency span: 0.38-0.04 = 0.34

would relative frequency span: 0.42-0.04 = 0.38

should relative frequency span: 0.19-0.03 = 0.16

so could and would are modals with largest relative frequency span

could modal:

austen-persuasion.txt uses it the most

whitman-leaves.txt uses it the least

would modal:

austen-emma.txt uses it the most

blake-poems.txt uses it the least

nltk.Text(gutenberg.words('austen-persuasion.txt')).concordance('could')

nltk.Text(gutenberg.words('whitman-leaves.txt')).concordance('could')

nltk.Text(gutenberg.words('austen-emma.txt')).concordance('would')

nltk.Text(gutenberg.words('blake-poems.txt')).concordance('would')

**Problem 2**. In the Inaugural corpus identify 10 most frequently used words longer than 7 characters. Which one of those has the largest number of synonyms? List all synonyms for those 10 words. Which one of those 10 words has the largest number of hyponyms? List all hyponyms of those 10 most frequently used “long” words.

from nltk.corpus import inaugural

fdist = FreqDist([word.lower() for word in inaugural.words() if len(word) > 7])

[str(i[0]) for i in fdist.most\_common(10)]

execfile("/Users/rpulekar/problem2\_synonyms.py")

**~~Problem 3.~~** ~~Create for us one graph displaying cumulative word length distribution for six different genres in Brown corpus. Create a tabular display of basic word statistics for all genres in Brown corpus. Include: average word length, average sentence length, number of concurrences in each genre, percentage of the text consumed by conditional words: would, could and should.~~

You literature for this assignment are chapters 1 and 2 of Natural Language Processing with Python book by Steven Bird et al.