### Assignment on Chapter 5a in “*Natural Language Processing with Python --- Analyzing Text with the Natural Language Toolkit” by* Steven Bird, Ewan Klein, and Edward Loper

You will find Chapter 5a posted at:

<http://www.nltk.org/book/ch05.html>

On the Canvas, a file named

nltkChapter5a\_template.py

is posted. You are requested to use this template in formulating your code solutions.

In all questions note how “The” and “the”, for instance, amount to the same word.

1. Train a unigram tagger on all of the sentences from the Brown corpus with the category *news* just as described in the “Unigram” Subsection of Section 5.1 of the textbook.
   1. Evaluate your tagger using “evaluate” function on all of the sentences from the Brown corpus with the category *lore.*
   2. How does this number compare to when this tagger is evaluate on all of the sentences from the Brown corpus with the category news.
   3. Provide the output of your tagger on the 200th sentence of the *lore* category of the Brown Corpus, (note how brown.sents(categories='lore')[199] produces the 200th sentence).
2. Write code to search the Brown Corpus for particular words and phrases according to tags, to answer the questions a-d. Report your findings separately for the following categories of Brown corpus: *humor, romance, government*.
   * 1. Produce an alphabetically sorted list of the distinct words tagged as JJ. Report the number of distinct words tagged as JJ and the first five words in the sorted list.
     2. Identify words that can be plural nouns or third person singular verbs (e.g. *deals*, *flies*). Sort these words alphabetically. Report the first 10 elements of the sorted list.
     3. Identify three-word prepositional phrases of the form IN + AT + NN (eg. at *the house*). Report the 3 most frequent three-word prepositional phrases (break ties alphabetically).
     4. What is the ratio of masculine pronouns

*himself, his, him , he, he's, he'd, he'll*

to feminine pronouns

*herself, her, hers, she, she's, she'd, she'll,*

where both are used with tag PP?

What are the differences that you observe with respect to different categories of Brown corpus.

1. Utilizing the complete Brown corpus report the following:
2. Analyze the tagged words to determine the number of distinct words that have exactly 5 possible tags. For example: The word *read* has 4 possible tags (NP, VB, VBN, VBD) found in the Brown corpus, whereas *debt* only has one possible tag (NN).
3. Determine which word(s) has the most distinct tags. Chose one of these word(s) and find sentences demonstrating the use of at least 5 distinct tags from the possible tags for the selected word.

Submit

* printed solutions to the assignment on a due date in the beginning of the class.
* file nltkChapter5a\_template.py (populated with your coding solutions) via Canvas 30 minutes before the class on a due date.