### Assignment (b) on Chapter 5 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 5 posted at:

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

On the Canvas, a file named

nltkChapter5b\_template.py

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

1. Write programs to process the Brown Corpus and find answers to the following questions:
   1. Which nouns are more common in their plural form, rather than their singular form (Use the parts of speech tags in the corpus to identify plural versus singular nouns and use nltk.WordNetLemmatizer() to get the singular form of a noun from its plural form). List the five most frequent nouns that feature this property.
   2. List the 5 most frequent tags in order of decreasing frequency. What do the tags represent?
   3. Which three tags precede nouns tagged with the 'NN' tag most commonly? What do these three tags represent? Report your findings separately for the following categories of Brown corpus: *humor, romance, government*.
2. In the “Combining Taggers” Subsection of Section 5.5 of the textbook, an example of a backoff tagger is provided. Extend that example by defining a TrigramTagger called t3 which backs off to t2. Train this tagger on all of the sentences from the Brown corpus with the category *news.* Then
3. evaluate your tagger using “evaluate” function on all of the sentences from the Brown corpus with the category *lore.* Report the number.How does this number compare to when this tagger is evaluated on all of the sentences from the Brown corpus with the category *news*.
4. 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). Would you tag this sentence in the same manner?
5. Compare the given TrigramTagger from the previous question with a TrigramTagger where no backoff is provided. Train this tagger on all of the sentences from the Brown corpus with the category *news.* Then evaluate your tagger using “evaluate” function on all of the sentences from the Brown corpus with the category *lore.* Report the numbers.Which tagger performs better? Why?
6. The majority of WordNet's senses are marked by four POS categories: noun, verb, adjective, and adverb. Determine the percentage of words from the WordNet corpus that have senses in more than one of these categories. For example, *type* has senses which connect to both “noun” and “verb” POS (positive case), whereas *typewriter* has only senses which connect to “noun” POS (negative case).

Submit

* printed solutions to the assignment on a due date in the beginning of the class.
* file nltkChapter5b\_template.py (populated with your coding solutions) via Canvas 30 minutes before the class on a due date.