CSC 206: NLP
Spring 2015

Name (Print): _____

Problem Set 4: Working with NLTK

Due: Friday, March 24 2015, before class

Please create a folder for this problem set, and save all your Python files in this folder as well as a (one) pdf document containing the answers to all the questions requiring a written answer.

Questions:

1. Describe the class of strings matched by the following regular expressions.
   1. `[a-zA-Z]+`
   2. `[A-Z][a-z]*`
   3. `p[aeiou]{,2}t`
   4. `\d+(\.\d+)?`
   5. `([^aeiou][aeiou][^aeiou])*`
   6. `\w+|[^\w\s]+`
2. Use the Porter Stemmer to normalize some tokenized text, calling the stemmer on each word. Do the same thing with the Lancaster Stemmer. Describe what differences you observe.

   See Section 3.6 for example of how to use the stemmers. As input text you can use any text from the NLTK corpora.

   Save your code in a Python file called `stemmers.py`. Make sure that this file contains all the Python you used to answer this question, so that when I run it, I can see exactly what data your written answer is based on.
3. Readability measures are used to score the reading difficulty of a text, for the purposes of selecting texts of appropriate difficulty for language learners. Let us define μw to be the average number of letters per word, and μs to be the average number of words per sentence, in a given text. The Automated Readability Index (ARI) of the text is defined to be: $4.71\mu w + 0.5\mu s - 21.43$. Compute the ARI score for various sections of the Brown Corpus, including section f (lore) and j (learned). Make use of the fact that `nltk.corpus.brown.words()` produces a sequence of words, while `nltk.corpus.brown.sents()` produces a sequence of sentences.
   1. Write a function called `brown_readability` that calculates and prints out the ARI for each genre/category in the Brown corpus.
   2. What do you observe? Do different genres have different readability scores? Would you have expected those differences? Explain.

Save your code in a Python file called `readability.py`.

Make sure that you use the function names and file names that are specified in the instructions.