CIS 4930 NLP // HW #10 // Spring 2018

Date Assigned: April 6, 2018
Date Due: April 13, 2018

Submission Format

You will submit a soft copy of your solution using e-Learning (http://elearning.ufl.edu) by the end of the day (23:59 / 11:59 PM) on the assigned date (April 13). Submit one file, **hw10.py**.

Assignment

At the top of every solution file you submit this semester include: your name, section number, the assignment number, and the date due. Complete the following exercises.

Exercises

1. [5.15]: Write programs to process the Brown Corpus and find answers to the following questions:

    a. Which nouns are more common in their plural form, rather than their singular form? (Only consider regular plurals, formed with the *-s* suffix.)
    b. Which word has the greatest number of distinct tags. What are they, and what do they represent?
    c. List tags in order of decreasing frequency. What do the 20 most frequent tags represent?
    d. Which tags are nouns most commonly found after? What do these tags represent?

2. [6.8]: Word features can be very useful for performing document classification, since the words that appear in a document give a strong indication about what its semantic content is. However, many words occur very infrequently, and some of the most informative words in a document may never have occurred in our training data. One solution is to make use of a **lexicon**, which describes how different words relate to one another. Using WordNet lexicon, augment the movie review document classifier presented in this chapter to use features that generalize the words that appear in a document, making it more likely that they will match words found in the training data.