Methods in Computational Linguistics 1
LING 78100
Fall 2010

Homework #5 - Corpus Statistics
Due at 11:50 on December 3

1. Short Problems -- Can be answered in the interpreter

A) [5] What is the most common part of speech tag in the brown corpus? What percentage of tags does it make up?
B) [10] What percentage of noun synsets have no hyponyms? You can get all noun synsets using wordnet.all_synsets('n')
C) [15] What is the average number of senses that each word in nltk.words.words() has in WordNet? len(w.synsets('dog','n')) will tell that the noun *dog* has seven senses.

2. Longer Programing Problems -- Processing many files.

A) [35] Write a program that reads a set of files one at a time and generates a vocabulary containing each unique lexical token in the files. After each file is written, print to the console (using the print command) the number of new words that were found. Report this information both as a raw number and a percentage. Finally, when the last file is written, write the vocabulary to a text file.
B) [20] Make a copy of the code you wrote in part A. This time, stem or lemmatize the text before adding the words to the vocabulary.
C) [5] Run this with the 3 Text normalizers. nltk.PorterStemmer() nltk.LancasterStemmer () nltk.WordNetLemmatizer()
D) [10] How does the stemmer impact the OOV rate? From this analysis which text normalizer would you prefer?