Programming Assignment 1

CS 4650/7650 Natural Language
Spring 2010

Due via e-mail before class on Thursday, February 16. Please start your file names with your name (e.g., david.novick.pa1.py).

Please do this assignment in teams of two. If possible, have the two team members come from different countries.

1. Write code to search the Brown Corpus for particular words and phrases according to tags, to answer the following questions:
   1. Produce an alphabetically sorted list of the distinct words tagged as MD.
   2. Identify words that can be plural nouns or third person singular verbs (e.g. deals, flies).
   3. Identify three-word prepositional phrases of the form IN + DET + NN (eg. in the lab).

   Turn in: Your code, plus use of the code showing answers to the questions.

2. Using the nltk file nltk/tag/sequential.py, write an n-gram back-off tagger that permits "anti-n-grams" such as ["the", "the"] to be specified when a tagger is initialized. An anti-n-gram is assigned a count of zero and is used to prevent back-off for this n-gram (e.g., to avoid estimating P(the | the) as just P(the)). Train and test your code on training and test sets from the Brown corpus. Compare evaluations of the normal nltk bigram tagger and your anti-n-gram bigram tagger.

   Turn in: Your code, the evaluation transcript, and a discussion of your results.