Homework 3
Due Tuesday, 02/19/08

Please turn in a print-out of your homework solutions at the beginning of class. If a program is required for a problem then you should provide sample input and output. (If the input is data from an on-line source then you may just indicate where the file can be found.) Programming solutions to problems should be placed in separate files whose names indicate the problem number (e.g. prob1.py). These separate files should be put into a folder whose name includes the assignment number and your name (e.g., Hmwk1-yourname). This folder should be dragged into the class dropoff folder at

`/common/cs/cs181/dropbox`

While I expect that I will be able to do most of the grading based on the papers that you hand in at the beginning of class, I would like to have access to the programs so that I can test them if necessary. I should be able to load your file into python and have it compile and run without error. All programming solutions should be fully documented and use good variable names.

Note that the files on our system can be accessed using ssh via the Mac server xserv.cs.pomona.edu or the linux server linus.cs.pomona.edu. Files and/or folders can be deposited into the dropbox remotely by using a program utilizing the sftp protocol.

Warning: There are several different versions of Bird et al floating around. From now on all assignments will be relative to the pdf edition of the entire book dated January 24, 2008, and available via the course web page.

1. Please do Problem 4.2 on page 39 of JM.
2. Please do Problem 4.3 on page 39 of JM.
3. Please do Problem 4.5 on page 39 of JM.
4. Please do Problem 3 on page 111 of Bird.
5. Section 4.4-4.5 of Bird discusses some taggers and using backoff to improve taggers. Use the techniques shown there (including backoff) to create the best tagger you can from the built-in taggers of nltk (many of which must be trained on already tagged data).

   Train your tagger on the (tagged) Brown corpus, group a. Then run your tagger on the untagged corpus, group a. Compare your results with the (officially) tagged data. Indicate what percent of the tags you got right. Discuss where most of your mistakes occurred and suggest methods that will improve its accuracy.