Homework 1
Due Tuesday, 02/05/08

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.

1. Exercise 1c and 1e in the exercises in section 2.7.3, pg 64 of Bird et al.
2. Exercise 2.2b in the exercises in section 2.7.3, pg 64 of Bird et al.
3. Exercise 4 in the exercises in section 2.7.3, pg 65 of Bird et al.
4. Exercise 2.1cdf in the exercises at the end of Chapter 2, pg 29 of Jurafsky and Martin. (But use python, not perl! Use "raw" strings to avoid problems with backslashes – see Bird, pg. 60, for details)
5. Exercise 2.4 in the exercises at the end of Chapter 2, pg 29 of Jurafsky and Martin.
6. The file wsj15-18.txt is available on the CS computer system at `/common/cs/cs181/data/wsj15-18`. This comprises over a megabyte of text from the Wall Street Journal. Please answer the following questions by writing python programs and utilizing the nltk library. (Be sure to turn in your programs as well as the answers.)
   (a) Find all words that include all of the vowels (a, e, i, o, and u) in order. Words are strings separated by whitespace. In particular, hyphenated words count as a single word.
   (b) Separate the contents of the file into sentences and words.
      i. How many unique words are in the text?
      ii. How many sentences are in the text.
      iii. What is the average word-length of a sentence?
      iv. Print out the shortest and longest sentences and report their lengths.

   Hint: Test your programs on shorter excerpts from this text file before running it on the whole file.