L555: Programming for (Computational) Linguists

September 1, 2010
In-Class Exercise #2

1) Go to Oncourse and download the files *dates_in_may.txt* and *data.tgz* and move them to your Desktop. (NB: "Dates in May …" originally taken from http://itre.cis.upenn.edu/~myl/languagelog/archives/002122.html)
2) Open a terminal.
3) Create a directory called *mydata* and create two directories there (without going to mydata first), called *data_in* and *data_out*.
4) Go to *mydata* and copy *dates_in_may.txt* and *data.tgz* to *data_in*.
5) Unpack *data.tgz*.
6) Translate all upper case characters in *vm.pos* to X and store the resulting text in *data_out* in a file called *vm.X*.
7) How many lines are different between *vm.pos* and *vm.X* (use diff)? Does wc do the trick?
8) Make *dates_in_may.txt* all uppercase and store the resulting text in *data_out* in a file called *DATES.txt*.
9) Sort *vm.pos* and uniq it.
10) Make a frequency dictionary of words in *dates_of_may.txt*: Convert all non-alphabetical characters to linebreaks (and squeeze them), sort them, uniq them (with count), and store the result in *data_out* in file *dates.dict*.
11) Sort *dates.dict* in descending order of frequency (option for descending order : -r).
12) Create a list of bigrams (sequences of 2 words): Create a wordlist from *dates_in_may.txt* and store it in *dates.l1*, then create a new version of this file by tail +2 dates.l1 > dates.l2. Then combine the two files with the command paste (which takes two input files as arguments) and save the result in *dates.bigram*.