Linguistics 439/539
Fall '14

Homework #1

1. Exercise 2.3 from Manning & Schütze, but change the values from 0.8 to 0.7 and from 0.0003 to 0.0005 respectively.
2. Using the `austen.txt` file from the website, write a Matlab program that does the following:
   (a) reads in the file;
   (b) separates the text into words;
   (c) reports the total number of words;
   (d) converts everything to lowercase and strips punctuation;
   (e) computes the counts of all words and reports the words and counts for the 10 most frequent words;
   (f) gives a plot (as a line) of the counts for the 100 most frequent words.
3. **539 students only**: Exercise 2.1 from Manning & Schütze.

Things to remember:

1. This is due by email by the **beginning** of class on **Sept. 16**. (You can email it to me before then if you prefer.)
2. Code must be in the form of a working/runnable m-file that you email to me.
3. Prose questions can be answered as a commented section in an m-file or in a pdf file. Microsoft Word documents are *not* acceptable.
4. The prose can be no more than *three* double-spaced pages.
5. Keep in mind that there may be funny symbols here that you've never used before. Leave time to make sure you have them right.
6. Remember: nothing late. Do not wait until the last minute to do this.
7. You may certainly talk to each other about this and other assignments, but everyone must turn in their own work. (If you do talk to others, I need to see evidence that you are doing more than listening and writing down what others say.)