Linguistics 439/539
Fall '14

Homework #5

1. Using the first 100,000 words of the Brown corpus from the website, build a bigram model. Normalize only by eliminating punctuation and converting to lowercase. Smooth correctly using Good-Turing. Test this on the four sentences in these instructions and report the predicted probabilities for these four sentences.
2. Imagine you're talking to an intelligent person who knows nothing about NLP and explain how Baum-Welch works. Give the math and an explicit HMM and show how the algorithm would work on your HMM.
3. **539 students only**: Explain how the inside-outside algorithm works, how it differs from and is similar to the forward-backward algorithm (Baum-Welch).

Things to remember:

1. This is due by email by the **beginning** of class on **Nov. 11**. Nov. 11 is a holiday, so you must turn this in by email. *Since it's a holiday, I really encourage you to organize your time so you can turn this in the day before so you can enjoy the holiday.*
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.)