# Homework 4
# Due Tuesday, 02/26/08

1. There is only one problem this week. Implement an HMM part-of-speech tagger using the Viterbi algorithm. Train your tagger using the `treebank.train` file available from the homework web page.

   Use the file in the `treebank.test` file to test your tagger. Note that this file is fully tagged, so you will need to strip off the tags before using it as input to your tagger.

   (a) Compare the results of your tagger with the original tagging of the test file. What is the accuracy of your tagger on the test data?

   *Answers of 86 to 90% seemed to be correct. Some differences might have been based on what to do about capital letters. I suspect that the ideal is to leave in capital letters except for the first word in a sentence.*

   *An important issue was what to do about words that hadn't been seen before. Just setting the probability to zero was clearly not a good strategy. Using some variant of Good-Turing would likely be helpful, but also just giving new words a very small probability would help – though technically it would make probabilities add up to something greater than zero.*

   (b) How does this compare to the accuracy obtained by always choosing the most frequent tag for each word?

   *Here again there was some variance – probably caused mainly by treatment of capitalization. Scores seemed generally to be within the range from a bit less than 84% to 88%.*

   (c) Identify five errors in the automatically tagged data, and analyse them.

   *There were a variety of errors here. Some came from removing caps from proper nouns (leading them to be identified as adjectives) to mistaking a possessive to be a verb contraction.*

   (d) Run your tagger on the larger training file `treebank.train.large`. Run it on the same test data file as before. What is the accuracy now?

   *Results here tended to be in the range of 93 to 94%, which isn't bad, but is not as good as humans. Most likely estimators gave answers around 86%, though one person claimed he got 92%, which seems too high.*