Kyle Tunis
Algorithms for NLP
Assignment 5

1. Baseline: TAGGING ACCURACY BY TOKEN: 21510/25097 = 85.7% OOV TOKENS: 711/2292 = 31.0%
   HMM: TAGGING ACCURACY BY TOKEN: 22032/25097 = 87.8% OOV TOKENS: 665/2292 = 29.0%
2. The two most frequently misclassified types are PROPN and NOUN, with 703 and 702 misclassifications respectively. This is most likely because these types are often out of vocab. Of the misclassifications of these two types, more than half were OOV.
3. All of the sentences with a higher gold than viterbi probability have either unusual grammatical constructions, or typos. The joint probability is capable of recognizing these unusual constructions but the Viterbi algorithm has a difficult time generating them.
4. Baseline: TAGGING ACCURACY BY TOKEN: 21009/25097 = 83.7%
   HMM: TAGGING ACCURACY BY TOKEN: 21679/25097 = 86.4%
   Accuracies on the Penn-tagged set is slightly lower than on the universal-tagged set. This is likely because Penn tagging has more possible tags, making it more difficult to learn.
5. Training takes roughly the same time for both tagsets, but tagging is significantly faster for the universal tagset. Training takes longer than tagging for both tagsets.
6. Training complexity is roughly linear on the number of words, meaning that runtime would likely scale linearly as the size of the training set increases. Predicting a single sentence's tags, however, is roughly linear on the number of *words in the sentence*, with no bearing on the size of the training data. Therefore, looking at only the time needed to predict a single sentence (and by extension the whole test set) would not change as training size increases.