Assignment 4 heuristics

Nitin Mahtani

Note that states are the tags and the evidence is the word observed.

First I trained my model on the test data by doing some counting. I created 3 probability tables: Initial probabilities (initial probability of each state/tag), Transition probabilities (probability of $state_x$ given $state_{x-1}$) and emission probabilities (probability a word occurs given the tag).
I stored these as dictionaries to improve the runtime.

Then I began the testing process. I split the test data by question marks, exclamation marks and periods to get each sentence separately.
I used the viterbi algorithm to get the probabilities of the sequence of tags and ran viterbi on each sentence separately.

Then for the missing words and missing tag combinations I assigned an epsilon probability to those values. This epsilon was chosen to be very small and after investigating it is much smaller than the average probability for the values.