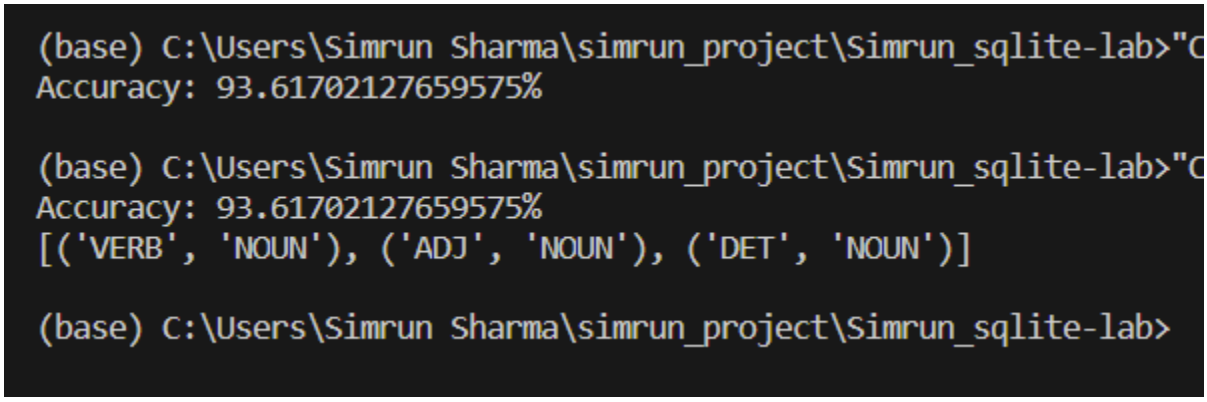
Part-of-speech tagging Homework Explanation



In this assignment we wanted to infer the hidden states for the sentence 10150-10152 of the Brown corpus. My POS tagger does not produce the correct tags however, it produces the right tags to a high accuracy (93.617 %).

I believe this accuracy doesn't reach 100 % for a couple of reasons. One reason being that the Viterbi algorithm is using test sentences it has never seen before, it might determine the word is UNK/OOV. That UNK/OOV highest probability state for the unknown word is a NOUN for me when it actually is a verb, adj, or det. Another reason is that Viterbi keeps track of all possible paths and their probabilities. However, because it has seen a training data that contains 10,000 words and not the whole english dictionary it might deem a certain path as having a higher probability simply because it a particular state sequence path is more prevalent in the Brown corpus but it might not be for the same if we were training with the whole English dictionary.