NLP Assignment 4

By Rohan Chhokra
2016080

This assignment asks us to implement the Viterbi algorithm to POS tag a test set given a manually annotated training set.

My implementation is as follows :

I'm reading the given training set line by line. Each line consists of a word and it's POS tag. Sentence stop is denoted by a fullstop. While reading the file, I'm reading the sentence start as <START> which has a START tag and the sentence end as <END> which has an END tag. Once the files have been made, I'm making a unigram dictionary which has the key values as a POS tag and it's value as it's frequency in the training corpus. Similarly I've made a nested dictionary which has a key as a POS tag and value as a dictionary have keys the following POS tag and value as frequency(basic POS bigram). Next I've stored all the words in the corpus in a dictionary with value as their frequency. If a word has less than or equal to 2 frequency I've changed it to <unk> and recalculated the dictionary with all less frequently occurring words as

<unk>. Lastly, to calculate the likelihood, I've made a nested dictionary with key value has a POS tag and value as another dictionary with key as the occuring word and value as it's frequency.

To implement Viterbi algorithm,I've made an adjacency matrix with dimensions - (number of POS tags,number of words) . I've done add-one smoothing for prior probability(count(tag1,tag2)+1/count(tag1)+number of tags) and for likelihood - (count(tag,word)+1)/(count(tag)+number of tags).