NLP Assignment-3 Report

Problem 1: The dataset here did not require any preprocessing. In the training set the sentences were given with each word followed by its POS tag separated by a tab. So, the sentences were separated using \n\n and then for each sentence the word and its corresponding tag was found by splitting on the basis of \t.

Now, I trained the model using the code written in Question1_Train.py and the testing code is Question1_Test.py.

Handling OOV words that is the words which are not present in the training set but are present in the test data:

Here using the method that the probability of the tag given an unknown word is very similar to the average of the probability of the singleton words given that tag in the training set.

So, I made a dictionary storing the average probabilities of such tags over singleton words.

The predicted tags are being written to a text file PredictedTags.txt in a format similar to the training set.

The model predicts the tags of the words of the sentence with good accuracy.