ASSIGNMENT REPORT

Name: Neethi Rajeeva Shetty

First I developed a **baseline model** that predicts the tag based on the frequency of the word-tag mapping. So in this method, I split the data to 80:20 for training and testing respectively. The frequency of each tag and frequency of each word to tag mapping was calculated . The unknown words seen in test data ,where tagged with the most frequently appeared tag. It's seen that the same sentences are repeated in the data set and is grouped together. Without shuffling the data in the data set, an accuracy of about 90 % was achieved, and with smoothing an accuracy of 94% was achieved.

Accuracy calc: number of correctly tagged words/total words in test set

Viterbi algorithm was implemented to solve the problem. The dataSet given was divided into 80:10 for training and testing respectively. To handle unknown words a new word <UNK> was added to word list and in observation likelihood matrix as well. <s> is considered as the previous tag for all the 1st words in the sentence. The transition matrix and observation matrix was calculated and laplace smoothing(add 0.5) was done on this. It was found that with add 1 smoothing there was less accuracy than add 0.5 frequency. Hence add 0.5 smoothing was used. With all these, without randomizing the dataSet an accuracy of 84% was achieved where as with shuffling, the accuracy of 91% is achieved.

Accuracy calc: number of correctly tagged words/total words in test set

To predict the tags for the test set posted for assignment2, all the data given previously was treated as train set. After training, the tags for the words in new test set was predicted and written to output file.