NLP Assignment - 3

Kaustav Vats | 2016048

Question 1

Preprocessing-
I removed last empty line from each of the data file.
For training I precomputed frequency of the transition probability and emission probabilities.
I consider “<Start>” as the start state and “<End>” as the last state.

To handle OOV words i did add-1 smoothing.
For Transition probability i added Vocab of Tags in denominator.
For Emission probability i added Vocab of words in denominator.

I also calculated accuracy for 90:10 split of the given data.

Accuracy	95.79

Question 2

Assumption-

1. Features take BIO tag also in consideration.
2. My understanding of the features.
    a. Feature 1 - Given BIO tag x, how many times a word w has occured as the start word of the sentence in the corpus.
    b. Feature 2 - Given BIO tag x, how many times a word w has occured as the last word of the sentence in the corpus.
    c. Feature 3 - Given BIO tag x, how many times a word w has previous tag as t in the whole corpus.
    d. Feature 4- Given BIO tag x, how many times a word w has next tag as the most occuring tag in the whole corpus.
    e. Feature 5- How many times a word had a bio tag x / Count of occurence of bio tag in whole corpus.

    Weight of each features would be, count of such feature/Word freq of a word given a bio tag x
    Features are one hot based on bio tag.

Accuracy for all bio tag on train.np and dev.np

BIO Tags	Train Accuracy	Test Accuracy
I-NP	94.918	87.793
B-NP	93.134	72.966
O	87.825	78.98