NLP HW2 Report

In order to complete the assignment I have followed the following actions at each step of the assignment.

1. Vocabulary Creation

In order to create the vocabulary I have taken the following steps:

- I have created the vocabulary by reading the training data line by line and storing the frequency of all the words in a dictionary.
- I used the split string function to get the length of each new line from the training data and kept only those words in the words frequency dictionary whose line length is greater than 1.
- Then I sorted the word frequency in descending order and finally wrote the vocabulary words with rank and frequency into the vocab.txt file.
- **What is the selected threshold for unknown words replacement?**
  I have chosen words with frequency **less than 3** for unknown words replacement.
- **What is the total size of your vocabulary and what is the total occurrences of the special token '< unk >' after replacement?**
  The total size of my vocabulary dictionary is **43193**. The total occurrence of '<unk>' token is **32537**.

2. Model Learning

In order to get the emission and transmission probabilities I have taken the following steps:

- I parsed the training data to get an entire list of individual sentences and their corresponding tags in a list of sentences and tags respectively.
- Then I created emission and transmission pairs dictionaries by parsing the above sentences and tags to get the number of emission and transmission pairs present in training data.
- Finally, I created the emission and transmission probabilities matrix by keeping " <-> " as the separator between the elements of the keys of both the emission and transmission probability dictionary.
- **How many transition and emission parameters are in your HMM?**
  Transmission Parameters Count: **23373**
  Emission Parameters Count: **1351**

3. Greedy Decoding With HMM

In order to get the POS tagging accuracy on dev data using the Greedy Decoding method, I have taken the following steps:

- Firstly I created initial probabilities of all the tags by dividing the frequency of each tag by the sum of the frequencies of all the tags.
- Then I split my dev data into a list of sentences and tags to pass the sentences into the greedy decoding function and get the calculated tags from it.
- In my greedy decoding function, I am calculating the emission and transmission probabilities of each word and tag and store the maximum probability index. Then finally storing the tags which had maximum probability.
- **What is the accuracy of the dev data?**
  My Greedy Decoding algorithm is giving **92.29455563904638** % accuracy on dev data.

3. Viterbi Decoding With HMM

In order to get the POS tagging accuracy on dev data using the Greedy Decoding method, I have taken the following steps:

- I have modified my greedy decoding algorithm to the predicted values from Viterbi decoding.
- Instead of directly keeping the tags with maximum probabilities, in the Viterbi algorithm I have calculated all possible probabilities combinations for each pair and then finally stored the index of the maximum of those.
- After getting the maximum probability index I added the corresponding tag to the result array.
- **What is the accuracy of the dev data?**
  My Greedy Decoding algorithm is giving **92.05091422455996** % accuracy on dev data.