CSCI 544 HOMEWORK 2

TASK 1:

What is the selected threshold for unknown word replacement?

I have used 3 as the threshold. The terms or words that have a frequency less than 3 will be considered as 'unk'

What is the total size of your vocabulary and what are the total occurrences of the special token '< unk >' after replacement?

Size of Vocabulary: 16920
Occurrences of special token <unk>: 32537

For task 1, I created a dictionary that will store the word as a key and will keep increasing its value by 1 as per the occurrence in the training data.

We are also considering a condition for threshold in which the words that have a total value less than 3 will be considered as <unk> token instead of their present word.

A new dictionary will be created to have <unk> token with its count as the first entry and the remaining words and their frequencies will be present in descending order. This dictionary will be created as a "vocab.txt" file.

TASK 2

How many transition and emission parameters in your HMM?

As I created nested dictionaries for both transition and emission parameters.
The length of the outer dictionary is:
Transition: 47 as it includes 'start' and 'fin'
Emission: 45

EXPLANATION

Transition matrix: Count of tags coming after the start or previous tag

Transition probability: Count of tags coming after the start or previous tag/Total Occurrence of start or previous tag

Emission matrix: In this nested dictionary is being used where each tag is key and then we have an inner dictionary for value where word and their frequencies are being counted for the corresponding tag according to the training data.

Emission probability: it is derived from an emission matrix, where each frequency is converted to probability for (tag, word) pair by diving the frequency with the number of occurrences of tags in the emission matrix

The dictionaries have been added in the form of JSON files.

TASK 3

What is the accuracy of the dev data?

I got an accuracy of around 93% for the greedy algorithm on the provided dev data.

EXPLANATION

Reading each word from the dev data provided, the function verifies the presence of the word

I am reading each line from the provided dev data that have a length greater than 1 and then picking up the word from the line.
Later, I am picking each tag from the 45 unique tags in which there will be 2 cases.
First, if we have the previous or start tag as 'start' for new sentences and find their transition probabilities with every 45 tags and find the maximum one and then use the found new tag(**current tag**) with max probability.

Find the emission probability with the current tag and word.

We will multiply both found emission and transition probability as total probability and the loop will go on for words to determine the tags.

These tags are appended to the list.

The tags will be then compared to dev data tags to find the accuracy.

TASK 4

What is the accuracy of the dev data?
I got an accuracy of around 86% for the Viterbi algorithm on the provided dev data.