NLP Homework 3 **Jayaprakash Jayakumar**
1209340128

Approach:

1. From the given training data, read the sentences word by word and tags associated for those words.
2. Created a dictionary to store the count of tags that occurs next to a tag. This count will be used to calculate the state transition probability between tags.
3. Created a dictionary to store the count for word's occurrences in a tag. This count will be used to calculate the observation probabilities.
4. Then, calculated the probabilities using the count values stored in the two dictionaries mentioned above. For handling unknown words, I have used Laplace Add one smoothing technique. Computed the Vocabulary, for state probabilities, Vocabulary is the total no of tags seen in the training set. For observation probability the vocabulary is the total no of distinct words seen in the training set. Using this smoothing is done while calculating the probabilities.
    a. aij = (count($q_t = s_j$, $q_{t-1} = s_i$) + 1)/(count($q_{t-1} = s_i$ + V)
    b. bj(k) = (Count($q_i = s_j$, $o_i = v_k$) + 1)/(Count($q_i = s_j$) + V)
5. Assigned a tag <unknown> and word <unknown> for each observation and calculated the probability for by treating the count as 0. Used this to estimate the probabilities for the unknown words that occur in the test dataset.
6. While calculating the state probability, also calculated the start probability from start state to each possible tags that can come for first word. And transition probability from last coming tag to end state. These two probabilities aij and bj(k) will be used as model lambda.
7. After calculating the probabilities, used the viterbi algorithm to maximize the probability to get the most likely sequence. This algorithm is used to estimate the tags for the test sequences.
8. Then based on the number of wrong assignment of tags and total no of words present in the test date, calculated the error rate.

Arbitrary sentence:

Sentence :
"It" "was" "the" "spartans" "," "who" "won" "the" "match" "."
Tag Assigned:
P V D N , W V D N .

Output:

```
Enter the training file path:
C:\Users\hp pc\Documents\Natural Language Processing\NLP Homework\3\entrain.txt
Enter the test file path:
C:\Users\hp pc\Documents\Natural Language Processing\NLP Homework\3\entest.txt
Training Completed...
```

NLP Homework 3 **Jayaprakash Jayakumar**
1209340128

```
Evaluation of test file started...
Total No of Words: 23949
Total No of wrongly tagged Words: 2350
Total No of correctly tagged Words: 21599
Error rate is : 0.0981251826798614
Error Percentage is : 9.81251826798614
Success Percentage is : 90.1874817320139
```

Code file is attached.
Unknown words are handled by Laplace Add one smoothing technique.