ASSIGNMENT 3

1. Preprocessing Steps:

- Split the corpus into Training and Testing Data.
- Now make sentences from the given training data set.
- To make sentence use split function on the given data of a file.
- Now calculate the following things and store in the pickles.
    - Tag-tag count
    - Word-tag count
    - First-tag count
    - Unigram-tag count

2. Assumptions:

- **Word-Tag Probability with add-1 smoothing:**

  P(word/tag)=(P(word->tag)+1)/(P(tag)+|V|)

- If tag is not present at the first position then its probability is 0.
- "." is consider as a tag in unigram tag list.
- "." Is not considered in word tag probability.
- Training data is divided into two parts 80% training and 20% testing.
- Sentence must end with "."

3. Accuracy:

The Accuracy of the model is 86.4% after taking 18 sentences from the testing data.

The accuracy is calculated by taking average of individual sentence accuracies.

4. Observations:

- For OOV words some time tags are correct.
- For Some words which are in corpus there parent tag may come not the child tag of the class. Example for nouns NN may come but NNS and similarly for other classes