Assignment -3

Task 1- HMM Model

1) Preprocessing

- Corpus is split into two parts where first part (larger) part is used for training of the model.

- Sentence tokenize is first used here to break the corpus in form sentences.

-Now regex is used to remove the "\n" and "\t" from the sentences.

- At last split is used to break the sentences in tokens.

-After that to apply HMM four basic things needed were extracted from the corpus that are-

1) Beginning word tags count.

2) Word+Tag count(for observation probability)

3) Tag+Tag count (for emission Probability)

4) Individual tag frequency.

Then Pickle file is created for all 4 of these.

2) Assumptions

- Smoothing is applied for the (bigram of tag + words) for those words which do not have appearance i.e (OOV). Smoothing is applied as –

Count of (word+tag) bigram + 1

Count of tag + |v|

Here |v| is unique tag counts.

-Empty text will not be assigned any "Tag"

-Test text will be entered in the same form as of the training data (without the tags).

-"." Is considered as a tag .

-Training Data is split into two parts pf (80% and 20%) . 80% for training the model, and 20% for testing the model.

3) <u>Accuracy</u>

- Accuracy Value is **85%** when averaged after taking 10+ sentences individual accuracies.

-It is calculated on the unseen test data drawn from the same corpus.

4) <u>Observation</u>

Sometimes the tag values change because of the bigram context i.e due to transitional probabilities. Even though tag class remains same i.e (class noun, or pronoun) but original tag differs.