CSPE73- Natural Language Processing

Assignment- 1

Due Date and Time: 6/10/2022, 11:59 PM **Marks:10**

In this assignment you are to implement **POS tagging** approach which is Viterbi algorithm for HMM -based approach. You will be using a bigram tag model.

Training:

I will be providing you a POS-tagged corpus (BERP corpus). Sentences in this dataset are arranged as one word per line and each word is assigned the tag in that line, both word and tag is separated by some blank space. A blank line precedes whenever a new sentence starts.

You need to assume that we do not encounter any new tags in the test dataset. But you have to consider the possibilities of encountering new words in the test data.

Decoding: (Viterbi algorithm)

Your model will read in sentences from a file which is similar to the train file except the tags (i.e. reading one word per line). And your output should be appropriate tags for each words. The output file format of your model should be same as the training file (i.e. word tag).

Evaluation

I will be providing an evaluation script for accuracy score and use the following command to use it

>> eval.py berp-gold.txt berp-predict.txt

Where berp-gold.txt contains the gold standard tags for respective words and berp-predict.txt contains your model output which is to be evaluated.

Address the following problems:

1. Calculate the maximum likelihood estimates for all the conditional probabilities required.
2. Handle unknown words
3. Perform smoothing (for transition probabilities)
4. Implement the Viterbi algorithm
5. Generate a confusion matrix for all the tag types.

I will be providing the test data few hours before the due time.

Deliverables

1. Code in one folder

2. A short report describing your choices and observation with respect to each of the 5 problems mentioned above (describing the purpose or functionality of each code file)
3. Include the results from test data in the report