Assignment 1: POS tagging

CS626: Speech and Natural Language Processing and the Web

Problem statement

- Implement a POS tagger in Python using approaches mentioned below
    - HMM
    - SVM
    - Bi-LSTM
- Input and output
    - Dataset: **Brown corpus**
    - Output: Accuracy (5-fold cross-validation), confusion matrix, per POS accuracy
- Create a document which reports the following for all three implementations
    - Compare the accuracy of all three models
    - Draw confusion matrix
    - Report per POS accuracy (accuracy for each tag)
    - Observe the strength and weaknesses of each model with respect to particular POSes
    - Perform detailed error analysis
    - Write a short paragraph on your learning.

Note

1. Use 5-fold cross-validation for reporting all accuracies
2. HMM, SVM and Bi-LSTM all need to be implemented from scratch

Dataset

- Brown corpus (Available in NLTK library) (http://www.nltk.org/nltk_data/)

Submission instructions

- The assignment is to be submitted in groups of 3 (Same group for every assignment and project)
- The submission link will be created on moodle to submit the assignment
- Only one person from the group with the lowest id is supposed to make the submission
- The name of the folder should be <id1_id2_id3>_Assignment1.zip
    - The uncompressed folder should contain three folders (HMM, SVM, Bi-LSTM, readme and a report in pdf format <id1_id2_id3_Assignment1>.pdf)
    - Each folder (for each approach) should contain their respective code files
    - The readme should contain details about the tools, versions, pre-requisites if any, and how to run the code for all three approaches.
    - The report should contain all things mentioned in the problem statement.
        - Accuracies, Per POS accuracies, confusion matrix, error analysis, strengths, and weaknesses of each model with respect to particular POSes, and a short paragraph on your learning.

Deadline

5 September 2020 (11:59 PM)

References

- https://www.nltk.org/book/ch05.html
- https://pythonprogramming.net/svm-in-python-machine-learning-tutorial/ (Follow the series, learn, don't copy the code)
- We shall check for code copying. Please be aware of neither copying codes from Git or amongst yourselves.