CS563-NLP

ASSIGNMENT-I : Part-of-Speech (PoS) tagging using HMM

(Read all the instructions carefully and adhere to them)

Date: Jan 25, 2023 **Deadline**: Feb 08, 2023

Scores: 20

Instructions:

1. Markings will be based on the correctness and soundness of the outputs.
2. Marks will be deducted in case of plagiarism.
3. Proper indentation and appropriate comments (if necessary) are mandatory.
4. You should zip all the required files and name the zip file as:

<roll_no1>_<roll_no2>_<roll_no3>_assignment_<#>.zip , eg. 2121cs33_2121cs34_2121cs35_assignment_01.zip.

5. Upload your assignment (the zip file) in the following link: https://www.dropbox.com/request/OZVRq6X65xXkpAGMv9NU

For any queries regarding this assignment contact: Aizan Zafar (aizanzafar@gmail.com), Ramakrishna Appicharla (ramakrishnaappicharla@gmail.com), Mamta (mamta20118@gmail.com) or Arpan Phukan (arpanphukan@gmail.com)

Problem Statement: Part-of-Speech (PoS) tagging assigns grammatical categories to every token in a sentence. In this assignment, you have to develop a PoS tagger using 2nd order Hidden Markov Model (HMM).

Dataset name: English Penn Treebank (PTB) corpus

Number of PoS tags: 36 (Alphabetical list of part-of-speech tags used in the Penn Treebank Project)

Link to download the dataset:

https://drive.google.com/file/d/1R1BLcghCh4j9Kl8_CR7MxZ4Wj57RiTxn/view?usp=share_link

- **Hidden Markov Model (HMM)**

  You have to implement HMM on your own. Do not use any existing libraries. Consider a bigram HMM model. Calculate the Emission and Transition Probability matrices. Use Viterbi decoding to obtain the best PoS sequence.

Evaluation:

1. Split the dataset in 80:20 ratio for train and test sets.
2. Compute and report the overall accuracy of HMM models on the test set.
3. Compute and report the class-wise accuracies.
4. Collapse all the 36 tags into 4 tags as follows: all the noun PoS tags to “N”; all the verb PoS tags to “V”; all the adjectives and adverbs to “A”, and the rest to “O”.
5. Repeat points 2 and 3 for (4).
6. Comment on the following: overall performance of 36-tag vs 4-tag model; if the overall performance of 4-tag is better than the 36-tag model, explain with intuition with respect to transition and emission probabilistic assumption why is such the case?

Note: *For the unseen word, consider the default PoS tag, which should be the most frequent tag in the entire dataset.*