CS 240: Lab 10
POS Tagging with HMM (Part II)

TAs: Deeptanshu Malu & Deevyanshu Malu

Instructions

- This lab will be **graded**.
- Please read the problem statement and submission guidelines carefully.
- For any doubts or questions, please contact either the TA assigned to your lab group or one of the two TAs involved in making the lab.
- The deadline for this lab is **Thursday, 3 April, 5 PM** but solutions till 5:30 PM will be accepted. No submissions will be accepted after 5:30 PM.
- The submissions will be checked for plagiarism, and any form of cheating will be penalized.

Problem Statement

You have to implement a POS tagger using the Hidden Markov Model (HMM) with the **Brown Corpus**. To download the Brown Corpus, the `nltk` library will be used.

The Brown Corpus contains 57340 sentences tagged with the POS tags according to the Universal POS tagset. The Universal POS tagset consists of 12 tags:

- `ADJ`: adjective
- `ADP`: adposition
- `ADV`: adverb
- `CONJ`: conjunction
- `DET`: determiner
- `NOUN`: noun
- `NUM`: numeral
- `PRON`: pronoun
- `PRT`: particle
- `VERB`: verb
- `.`: punctuation
- `X`: other

Note

This lab is **part II of a two-part lab**. In this part, you have to implement the Viterbi algorithm to find the POS tags for a given sentence.

Tasks to be Completed

Task 1. Implement the Viterbi algorithm to find the most probable sequence of tags for a given sentence. It is preferable to use the log probabilities to avoid underflow.

Task 2. Get the predicted tags for the test set in each iteration of the cross-validation. Concatenate the predicted tags for all the folds to get the final predicted tags. Do similarly for the actual tags. Finally give the classification report on the final predicted tags and the actual tags.

Task 3. Create a function that takes a sentence as input and returns the POS tags for the sentence using the Viterbi algorithm (Use the full dataset for making the transition and emission matrices).

Submission

- Submissions should be made on Moodle. Submit the Jupyter Notebook file renamed as `rollnumber1_rollnumber2.ipynb` (the ”b” in roll number should be in small case).
- Penalty will be imposed on wrong file naming.
- The hard deadline for submission is 5:30 pm. No submission after that will be evaluated.
- Only one person per team should submit their solution.