CS 240: Lab 9
POS Tagging with HMM (Part I)

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, 27 March, 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 I of a two-part lab**. In this part, you have to implement the training part of the HMM model, i.e., creating the transition matrix and the emission matrix.

Tasks to be Completed

Task 1. Make 5 folds for K-fold cross-validation.

- Take 4 folds at a time to create the training set and the remaining 1 fold as the test set.
- Repeat this process 5 times to get 5 different training and test sets.

- For each fold, create a transition matrix and an emission matrix using the training set.
- Use **Add-one (Laplace with $\alpha = 1$) smoothing** while calculating the transition and emission probabilities. Add-one smoothing is used to handle the cases when the probability is zero.

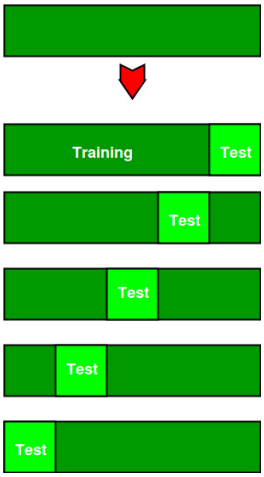

Figure 1: 5-fold cross-validation

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.