NLP Assignment 2: Task 1: POS Tagging

--Utsav Basu (20CS30057)

Problem Statement

Given the POS tags in the train dataset, we need to calculate the transition and emission probabilities of each word and tag, and create the Viterbi matrix from scratch using these values. The classification reports on both the train and test data should be shown.

Dataset

We are using the UD-English-GUM dataset. This dataset consists of many sentences already tokenized into words. Each word is annotated with its POS tag and its head word in the dependency graph, along with the type of dependency relation. A fictional root word is added as the root of the dependency graph. Each sentence starts with a unique `# sent_id` and a `# text` describing the text of the sentence.

Methods and Results

POS tagging is done by the Viterbi algorithm. Emission/transition probabilities were learnt from the train set and then applied to the test set. Since probabilities can be quite small, their -ve log was taken to avoid overflow error during calculations. The algorithm was also modified accordingly to account for the -ve log. Smoothing has been during all probability calculations to avoid zeroing out the full result.

The results obtained are as follows:

Train:

Accuracy: 0.9343706033269883
Recall: 0.9201414275358814
Precision: 0.8870908691830516
F1: 0.8996635609098782

Test:

Accuracy: 0.7842056932966024
Recall: 0.8029331104985741
Precision: 0.7574504988507769
F1: 0.7591166166687221