Assignment 1: Sequence Labelling Task using Viterbi Algorithm

Introduction

To implement a Part-of-Speech (POS) tagger using the Viterbi Algorithm.

- Training the tagger on the Brown Corpus to estimate transition probabilities and emission probabilities.
- Implementing the Viterbi Algorithm for sequence labelling.
- Reading input sentences from a file, tagging them, and writing the tagged output to another file.

Implementation Details

The implementation can be broken down into several essential components:

Data Preparation

- Loading the Brown Corpus, made available through the NLTK library, to facilitate training and probability estimation.
- Calculating transition probabilities, emission probabilities, and initial probabilities based on the tagged sentences within the corpus.

Viterbi Algorithm

- Viterbi Algorithm to determine the most appropriate tag sequence for a given sentence.
- Utilising dynamic programming techniques, including forward pass and backtracking, to compute the most probable sequence of POS tags.

Input and Output Handling

- Reading input sentences from a designated input file, with each sentence on a separate line.
- Applying the Viterbi Algorithm to tag the input sentences.
- Writing the tagged sentences to an output file, adhering to the specified format.

Usage Instructions

1. NLTK library is correctly installed and Brown Corpus is downloaded.
2. Input file (e.g., input.txt) containing one sentence per line.
3. The code will proceed to process the sentences, apply tagging, and generate tagged output in the specified output file (e.g., output.txt).