**Lab 5**

**Lab Aim**: Perform the experiment to calculate emission and transition matrix for tagging Parts of Speech (PoS) using Hidden Markov Model and Viterbi decoding.(Based on Virtual Lab https://nlp-iiith.vlabs.ac.in/)

**Objective**:

* The objective of the experiment is to calculate emission and transition matrix which will be helpful for tagging Parts of Speech using Hidden Markov Model.
* The objective of this experiment is to find POS tags of words in a sentence using Viterbi decoding.

**Dataset or Resources Used** - Virtual Lab by The International Institute of Information Technology Hyderabad (https://nlp-iiith.vlabs.ac.in)

**Procedure Followed** -

**POS Tagging – Hidden Markov Models**:

1. Select the corpus.
2. Fill the column with the probability of possible POS tags given the word (i.e. form the viterbi matrix by filling column for each observation). Answers submitted are rounded off to 3 digits after decimal and are than checked.
3. Check the column.
4. Repeat steps 2 and 3 until all words of a sentence are covered.
5. At last check the POS tag for each word obtained from backtracking.

**POS Tagging – Viterbi Decoding:**

1. Select the corpus.
2. Fill the column with the probability of possible POS tags given the word (i.e. form the viterbi matrix by filling column for each observation). Answers submitted are rounded off to 3 digits after decimal and are than checked.
3. Check the column.
4. Repeat steps 2 and 3 until all words of a sentence are covered.
5. At last check the POS tag for each word obtained from backtracking.

**Date of Performance** - 29 January, 2025

**Theory:**

**POS tagging**

POS tagging or part-of-speech tagging is the procedure of assigning a grammatical category like noun, verb, adjective etc. to a word. In this process both the lexical information and the context play an important role as the same lexical form can behave differently in a different context.

**POS Tagging - Hidden Markov Model**

A Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states.In a regular Markov model (Markov Model (Ref: http://en.wikipedia.org/wiki/Markov\_model)), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible.

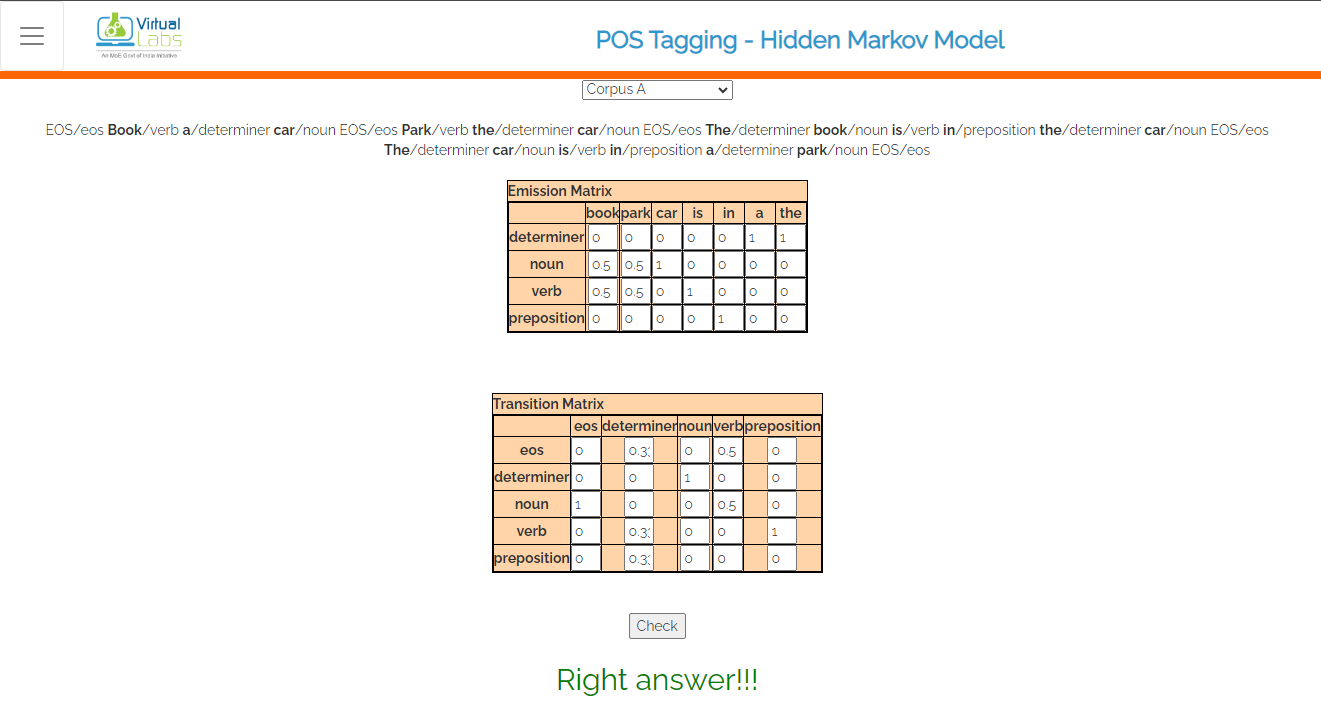
**POS Tagging - Viterbi Decoding**

Viterbi Decoding is based on dynamic programming. This algorithm takes emission and transmission matrix as the input. Emission matrix gives us information about proabities of a POS tag for a given word and transmission matrix gives the probability of transition from one POS tag to another POS tag. It observes sequence of words and returns the state sequences of POS tags along with its probability.

**Code & Output -**

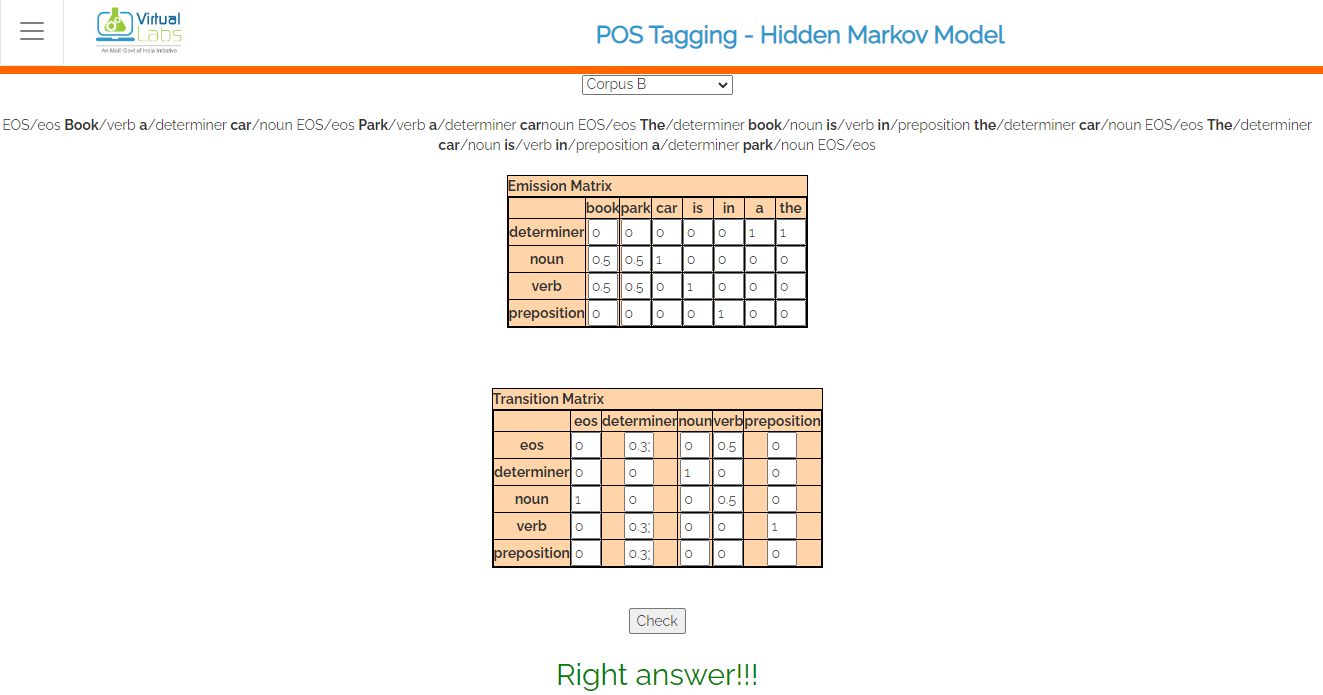
**POS Tagging – Hidden Markov Models**

**CORPUS A**

****

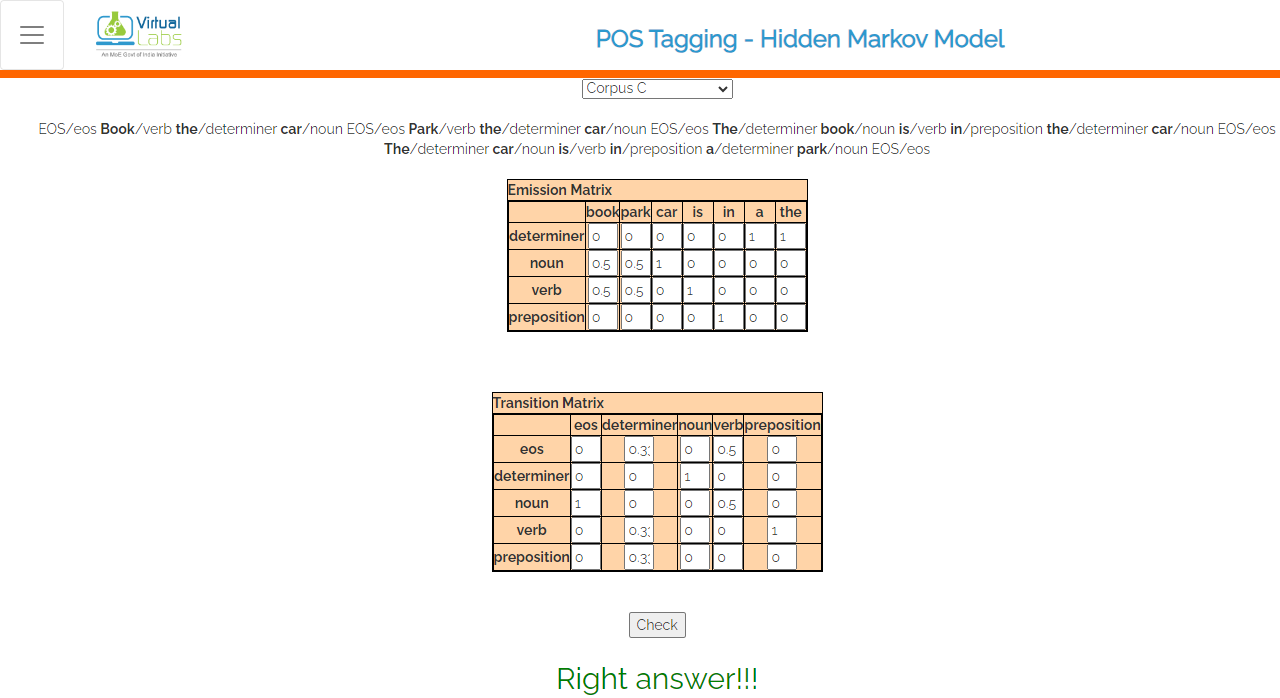
**Correct Emission and Transition Matrix for Corpus A**

**CORPUS B**

****

**Correct Emission and Transition Matrix for Corpus B**

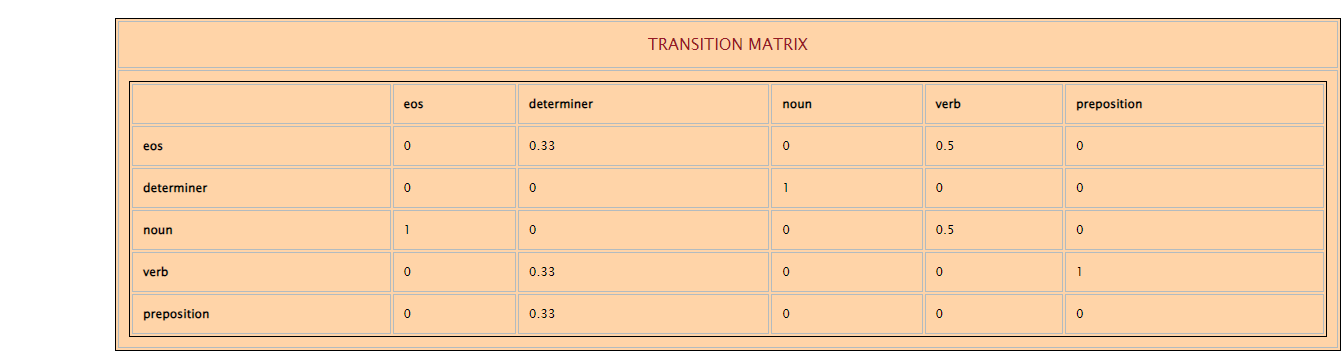
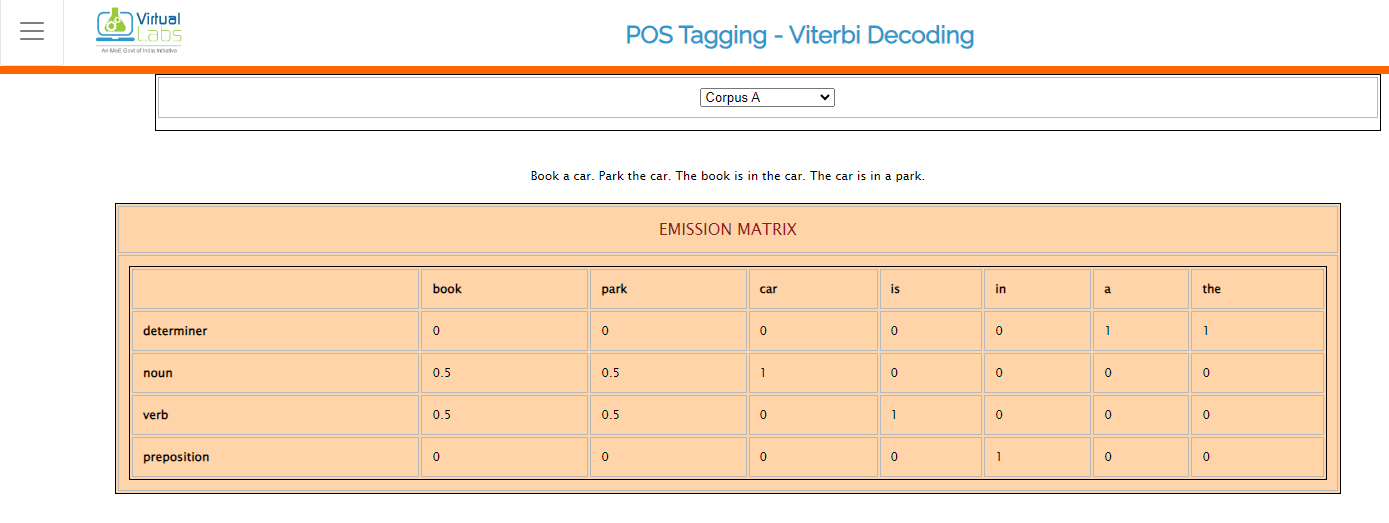
**CORPUS C**

****

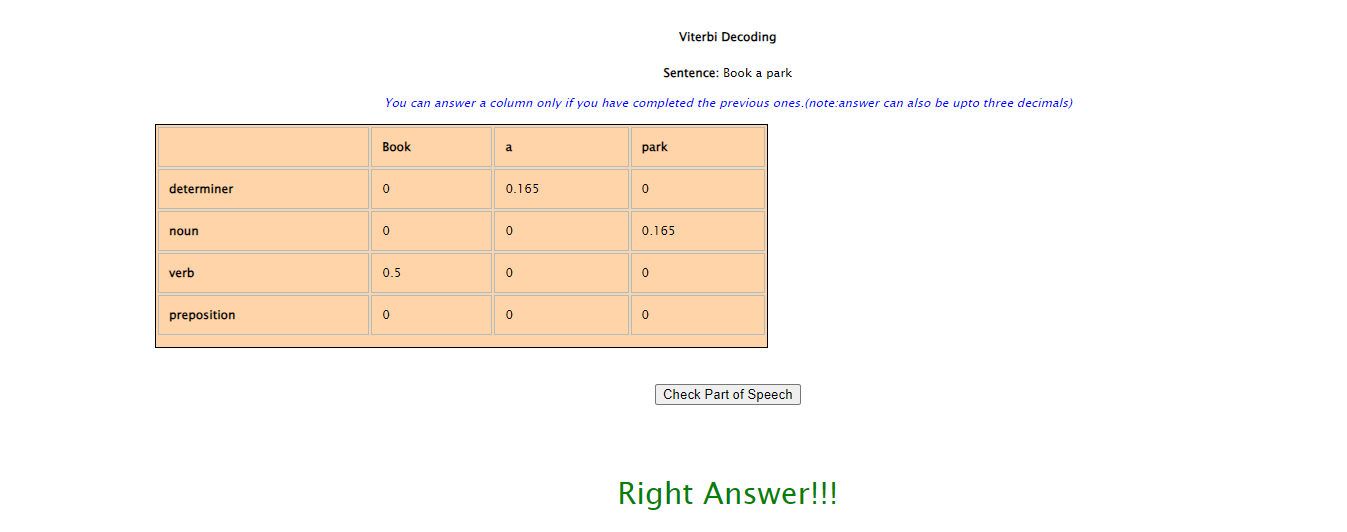
**Correct Emission and Transition Matrix for Corpus C**

**Code-B:**

**POS Tagging – Viterbi Decoding**

****

**Emission and Transition Matrix for Corpus A**

**** 

**Correct Viterbi Decoding for Corpus A**

**Viva Questions**

**1. What is part-of-speech tagging, and why is it important in natural language processing (NLP)?**

Ans: Part-of-speech tagging assigns grammatical categories (such as nouns, verbs, adjectives) to words in a sentence. It’s crucial for understanding the syntactic structure of text, disambiguating word meanings, and enabling downstream NLP tasks.

**2. What are emission and transition probabilities in the context of Hidden Markov Models (HMMs) for POS tagging?**

Ans: Emission probabilities: These represent the likelihood of observing a specific word given a particular POS tag. They help determine the most likely POS tag for a word. Transition probabilities: These capture the probability of transitioning from one POS tag to another in a sequence. They guide the overall tagging process.

**3.** **Explain the Viterbi decoding algorithm and its role in POS tagging.**

Ans: The Viterbi algorithm is used to find the most likely sequence of POS tags given a sequence of words. It combines emission and transition probabilities to compute the optimal tag sequence.By considering the probabilities of different tag sequences, Viterbi decoding helps improve the accuracy of POS tagging.

**Conclusion**

Part-of-speech (POS) tagging is a fundamental task in natural language processing (NLP) that involves assigning grammatical categories (such as nouns, verbs, adjectives) to individual words in a sentence. While manual rule-based approaches are limited by the patterns identified by human rule crafters, machine learning techniques offer a more robust solution. By training models on annotated corpora, we can automatically learn patterns from data and create accurate POS taggers. The Viterbi decoding algorithm, which utilizes transition and emission probabilities, is commonly used to find the most likely sequence of POS tags for a given sentence. Understanding these techniques allows us to analyze and process text effectively in various NLP applications.