**Terna Engineering College**
**Computer Engineering Department**
**Program: Sem VIII**

**Course: Natural Language Processing**

# Experiment No. 3

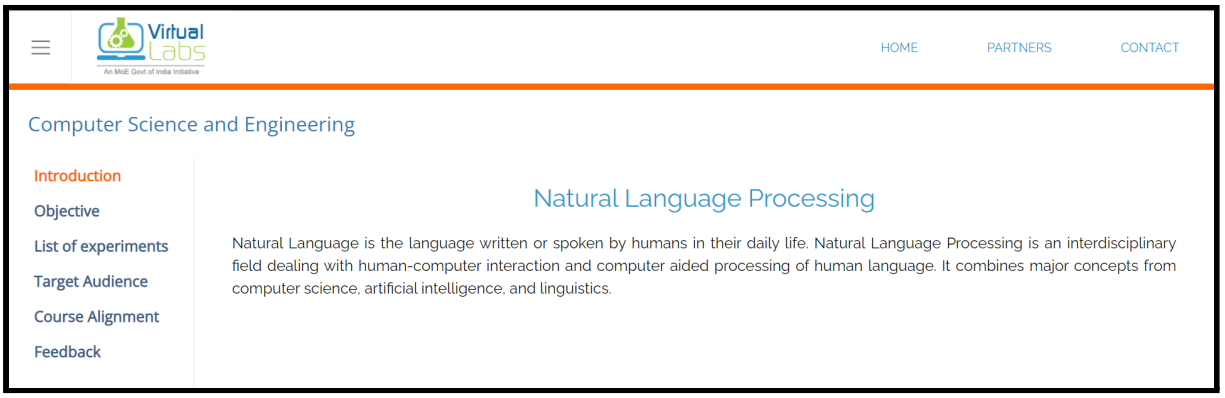
**A.1 Aim:** Perform and analyse an n-gram modelling for corpuses using Virtual Lab.

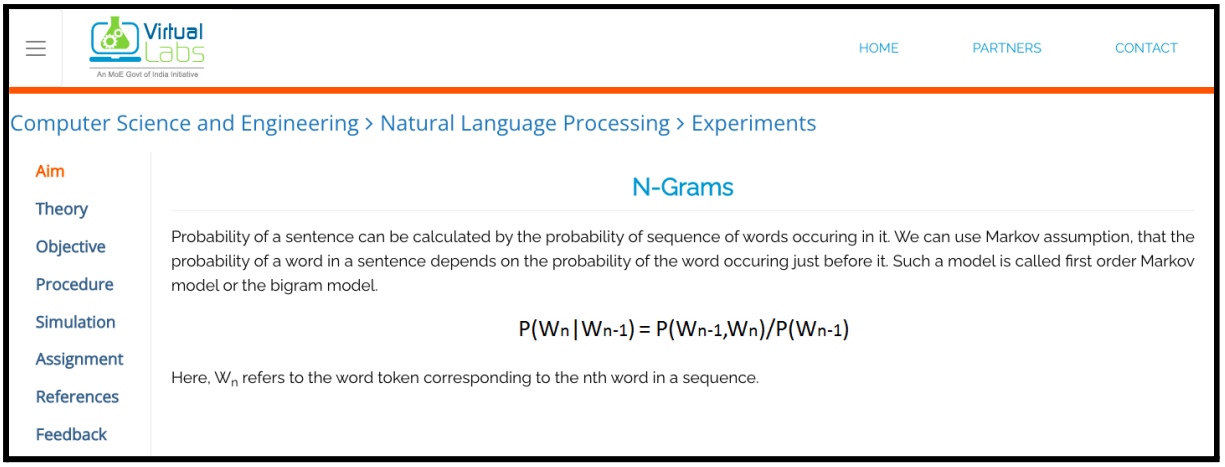
---

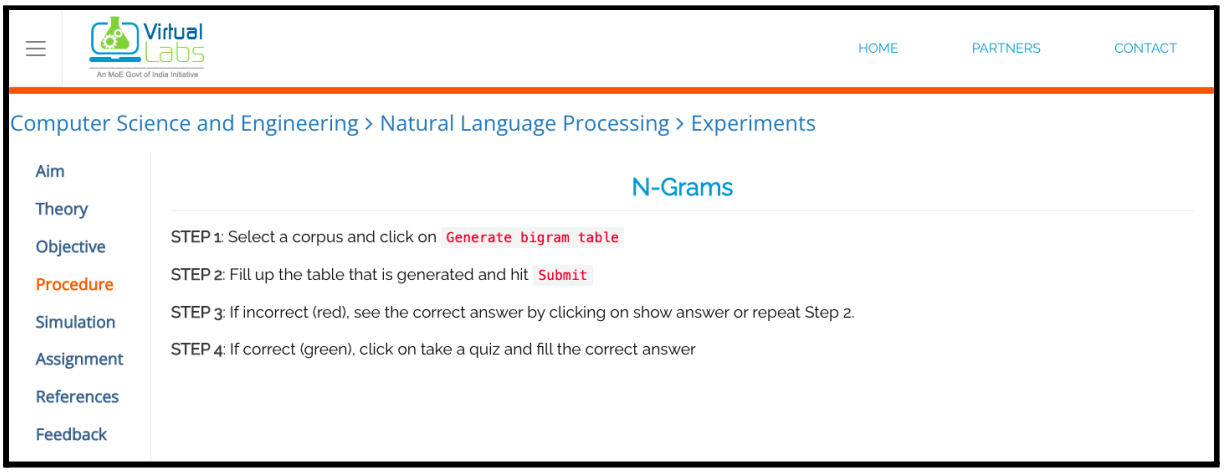
**PART B**
**(PART B: TO BE COMPLETED BY STUDENTS)**

| | |
|---|---|
| **Roll No.** 50 | **Name:** AMEY THAKUR |
| **Class:** BE COMPS B | **Batch:** B3 |
| **Date of Experiment:** 07/02/2022 | **Date of Submission:** 07/02/2022 |
| **Grade:** | |

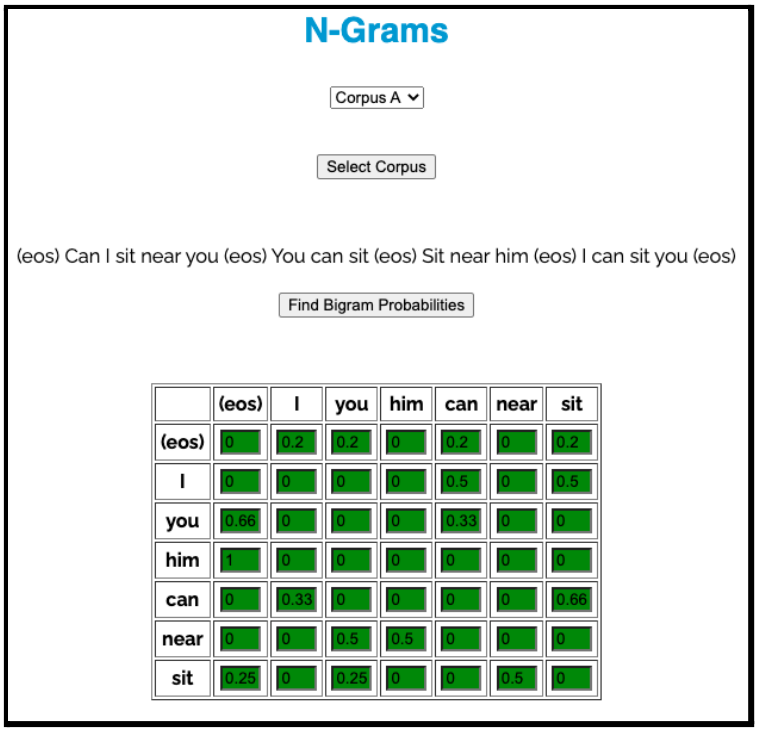
## B.1 Virtual Lab (Input & Output):



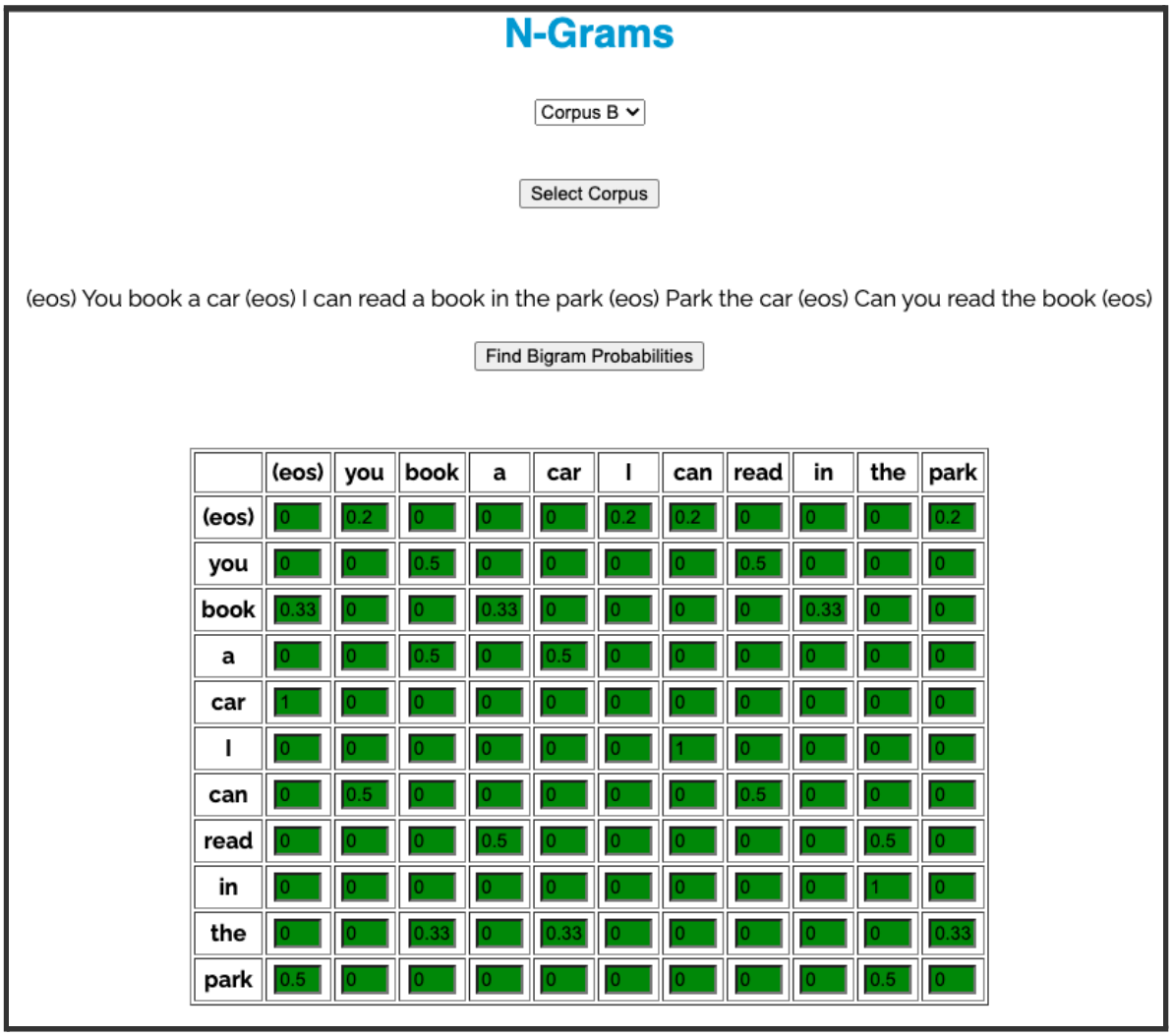




## CORPUS A:




| | (eos) | I | you | him | can | near | sit |
|---|---|---|---|---|---|---|---|
| (eos) | 0 | 0.2 | 0.2 | 0 | 0.2 | 0 | 0.2 |
| I | 0 | 0 | 0 | 0 | 0.5 | 0 | 0.5 |
| you | 0.66 | 0 | 0 | 0 | 0.33 | 0 | 0 |
| him | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| can | 0 | 0.33 | 0 | 0 | 0 | 0 | 0.66 |
| near | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0 |
| sit | 0.25 | 0 | 0.25 | 0 | 0 | 0.5 | 0 |

**CORPUS B:**




| | (eos) | you | book | a | car | I | can | read | in | the | park |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (eos) | 0 | 0.2 | 0 | 0 | 0 | 0.2 | 0.2 | 0 | 0 | 0 | 0.2 |
| you | 0 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0 |
| book | 0.33 | 0 | 0 | 0.33 | 0 | 0 | 0 | 0 | 0.33 | 0 | 0 |
| a | 0 | 0 | 0.5 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| car | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| can | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0 |
| read | 0 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0 |
| in | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| the | 0 | 0 | 0.33 | 0 | 0.33 | 0 | 0 | 0 | 0 | 0 | 0.33 |
| park | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0 |

**B.2 Observations & Learning:**

- Text N-grams are commonly employed in text mining and natural language processing.
- They're essentially a group of co-occurring words inside a particular window, and when computing the n-grams, you usually move one word ahead (but in more complex cases, you can move X words forward).

**B.3 Conclusion:**

We have successfully performed and analysed an n-gram modelling for corpuses using Virtual Lab.

**B.4 Questions of Curiosity:**

Q1. What is N-gram? What is its purpose and need?
ANS:

- Text N-grams are commonly employed in text mining and natural language processing. They're essentially a group of co-occurring words inside a particular window, and when computing the n-grams, you usually move one word ahead (but in more complex cases, you can move X words forward).
- N-grams are utilised for a wide range of purposes. When creating a language model, for example, n-grams are used to create not just unigram models but also bigram and trigram models. Google and Microsoft have created web-scale n-gram models that may be used for spelling correction, word breaking, and text summarising.

Q2. Give an example of the application of the N-gram used in NLP from a recent research paper.
ANS:

- Text Categorization based on N-grams
- N-gram Machine Classification of Sentiment Reviews
- N-gram Method in Large-Scale Clustering of DNA Texts
- A Compromise between N-gram Length and Classifier Characteristics for Protein Classification
- N-Gram characterization of genomic islands in bacterial genomes

Q3. Find bigram probabilities for given sentences (CORPUS).

1. can one play on ground
2. only work no play
3. one is on ground

ANS:

| | can | one | play | on | ground | only | work | no | is |
|---|---|---|---|---|---|---|---|---|---|
| can | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| one | 0 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0.5 |
| play | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| on | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| ground | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| only | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| work | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| no | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| is | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |