**Lab Practical and date – Practical 6, Tuesday 3rd May 2022**

**Name and Roll Number- Naman Thaker (20BCE529)**

**Practical Objective- N Grams , N Grams Smoothing**

1. **N grams**

**The chance of a sentence occurring in a particular sequence**

**of words could well be determined. We may utilize the Markov**

**assumption, which states that the chance of a word appearing in a**

**phrase is proportional to the probability of the word appearing**

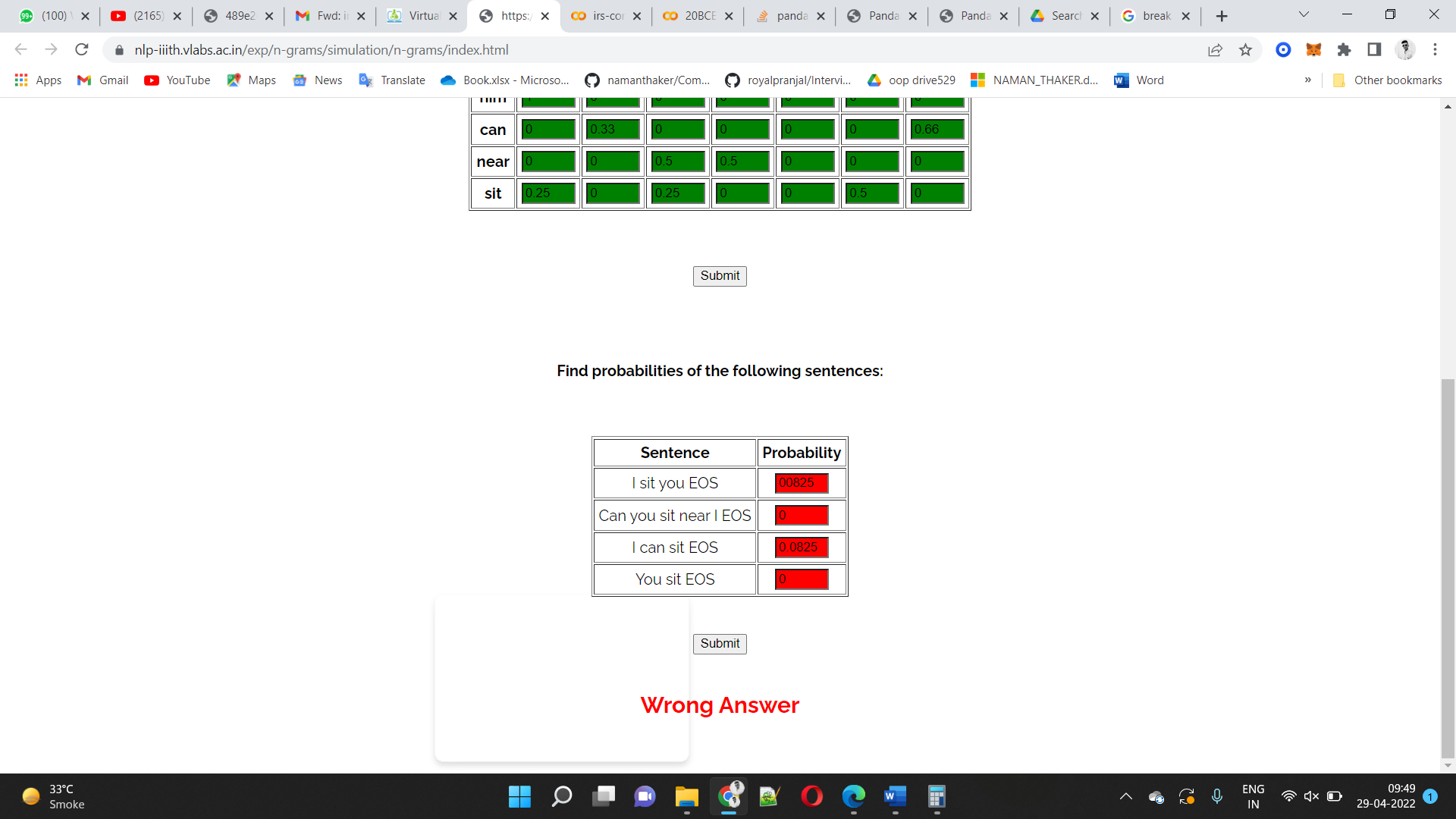
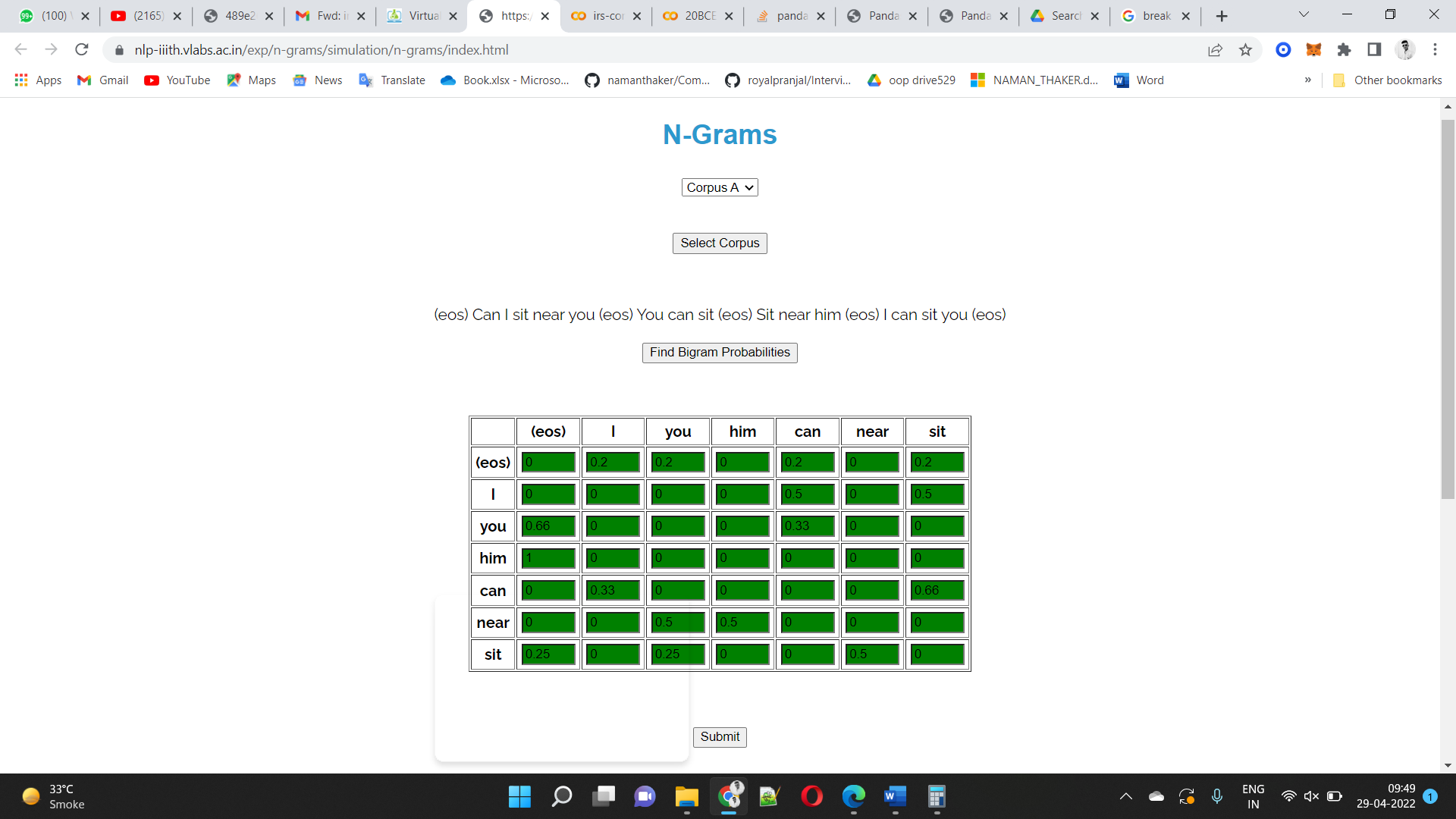
**immediately before it. The first order Markov model, also known as**

**the bigram model, is one such model.**

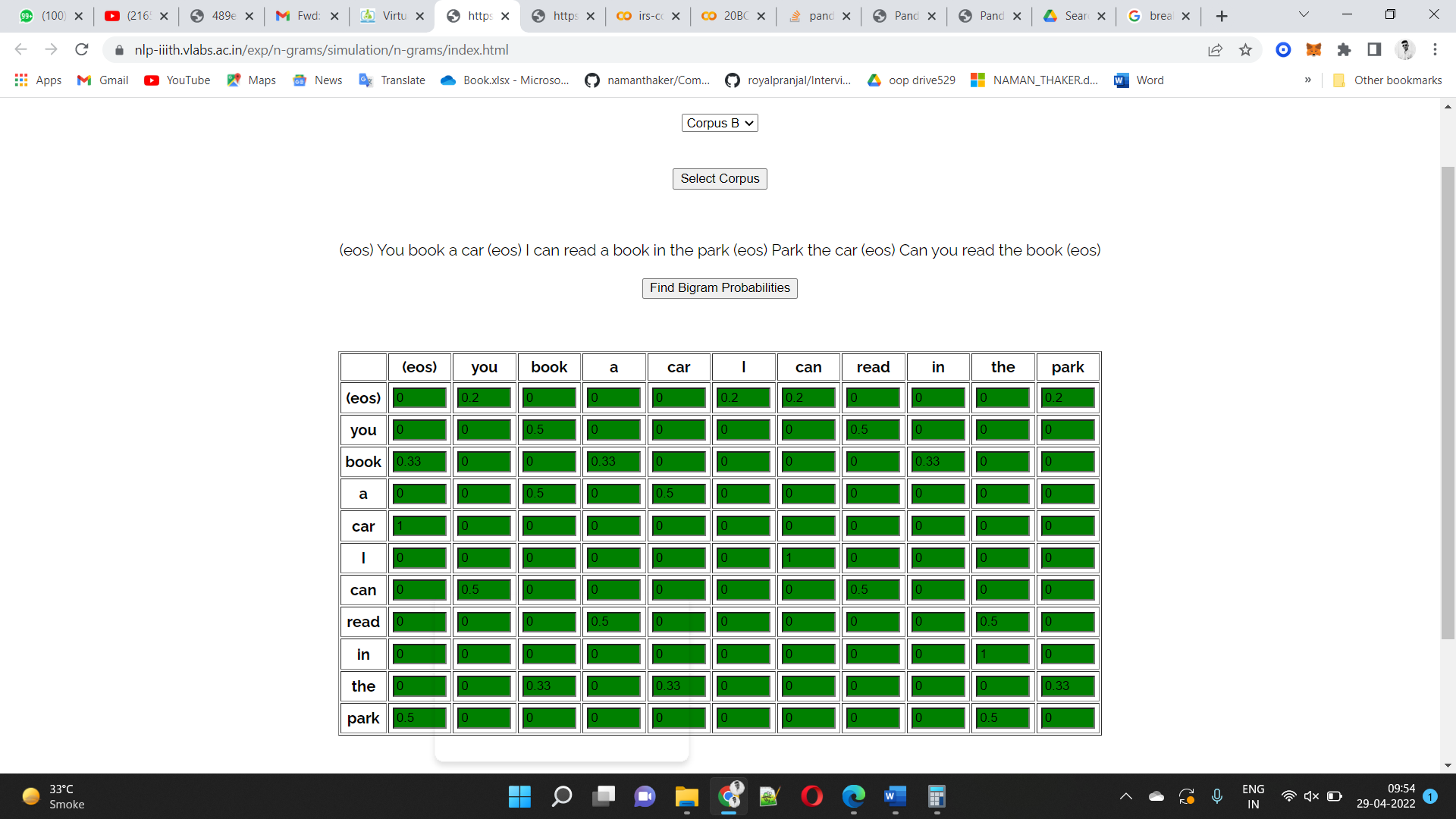
**OBJECTIVE: To calculate bigrams from a given corpus and**

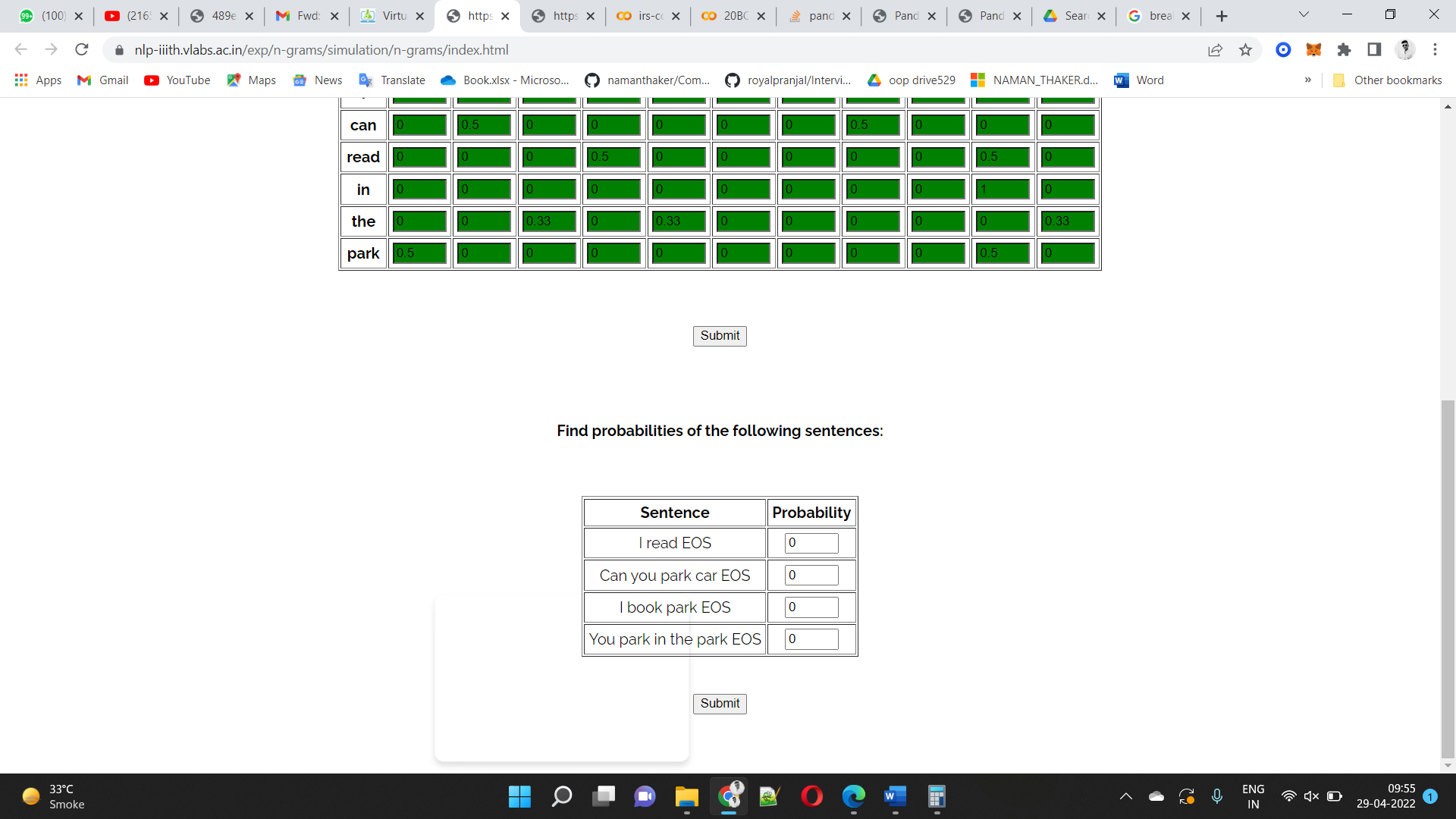
**calculate probability of a sentence.**

Corpus 1



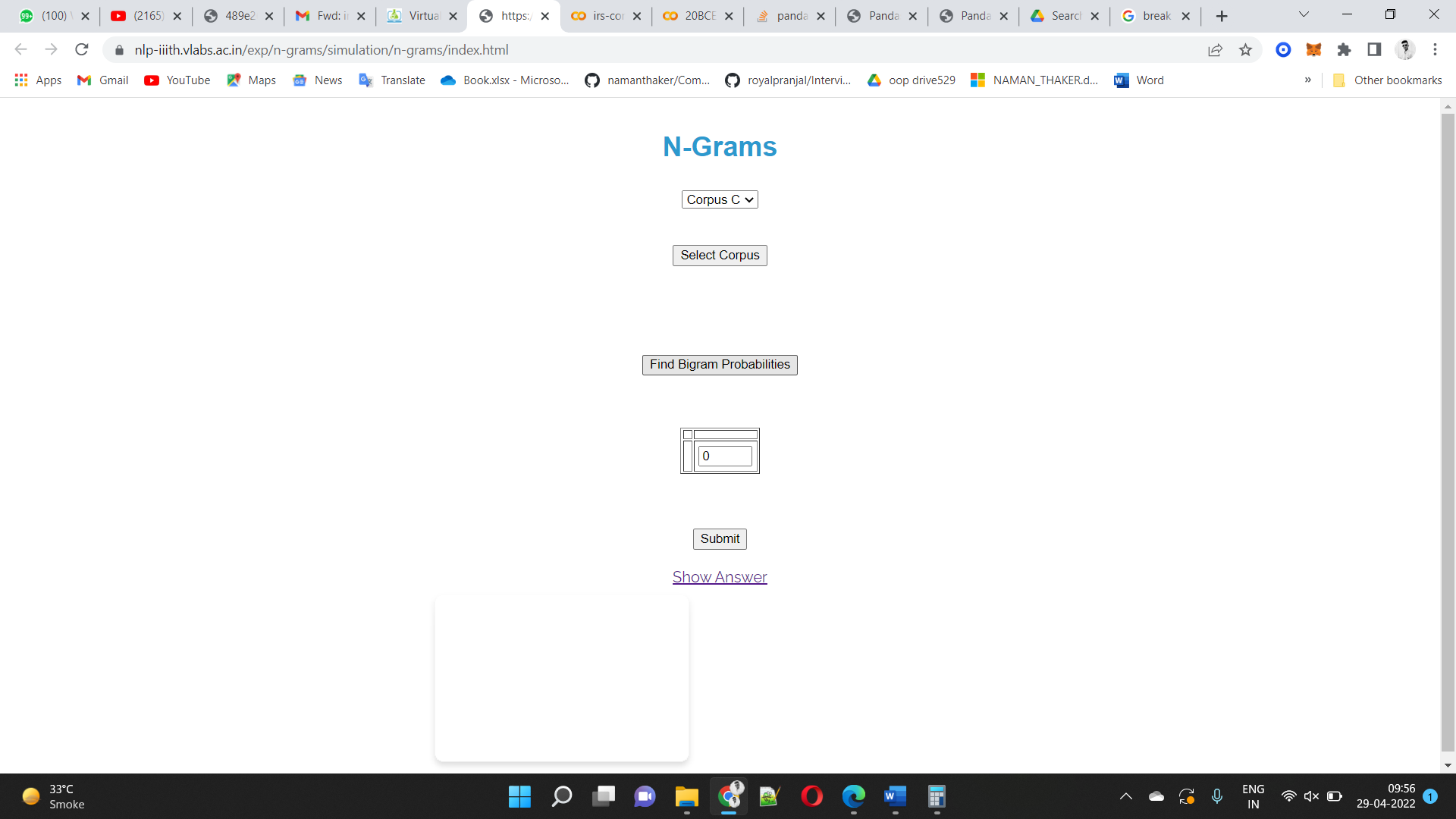
Corpus 2





Corpus 3

No corpus given



Assignment Question:

A trigram is a second-order Markov model. Derive the formula to calculate trigram probability. Next, calculate the trigram probabilities for the given corpus.

**(eos) Can I sit near you (eos) You can sit (eos) Sit near him (eos) I can sit you (eos)**

**Formula :**

**P(w1,w2,w3,…wn) = P(W1) \* P(W2|W1) \* P(W3|W1W2) \* P(W4|W2W3) \*…… \* P(Wn|Wn-2W2-1)**

Non zero Trigram probabilities

|  |  |
| --- | --- |
| **Trigram** | **count** |
| eos | 1 |
| sit eos sit | 1 |
| can sit you | 1 |
| I can sit | 1 |
| eos I can | 1 |
| him eos I | 1 |
| near him eos | 1 |
| sit near him | 1 |
| eos sit near | 1 |
| can sit eos | 1 |
| eos Can | 1 |
| You can sit | 1 |
| eos You can | 1 |
| you eos You | 1 |
| near you eos | 1 |
| sit near you | 1 |
| I sit near | 1 |
| Can I sit | 1 |
| eos Can I | 1 |
| sit you eos | 1 |

1. **N Gram Smoothing**

AIM: Standard N-gram models have one key flaw: they must be

trained from some corpus, and because every training corpus is

limited, some perfectly good N-grams are guaranteed to be

missed. The bigram matrix for any given training corpus is sparse,

as can be shown. There are a lot of scenarios with zero probability

bigrams that should have non-zero probability. This approach

tends to undervalue the likelihood of strings that did not appear

close in their training corpus.

There are various ways that can be utilized to give these

probability bigrams; a non-zero probability. Smoothing is the

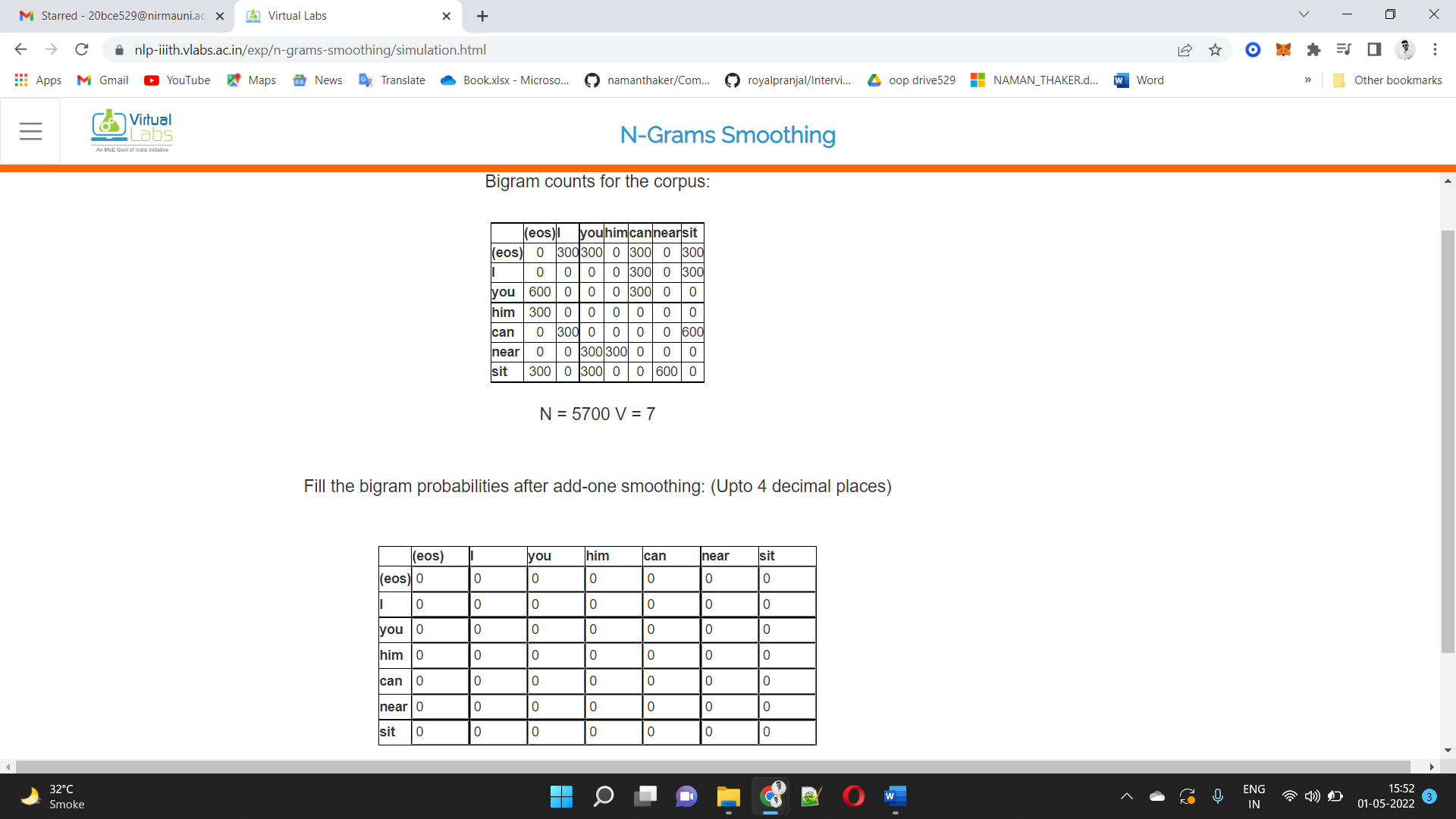
process of reevaluating and assigning non-zero values to some of

the zero-probability and low-probability N-grams.

OBJECTIVE: To apply add-one smoothing on sparse bigram table.

Corpus A

Question :



Output :

