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

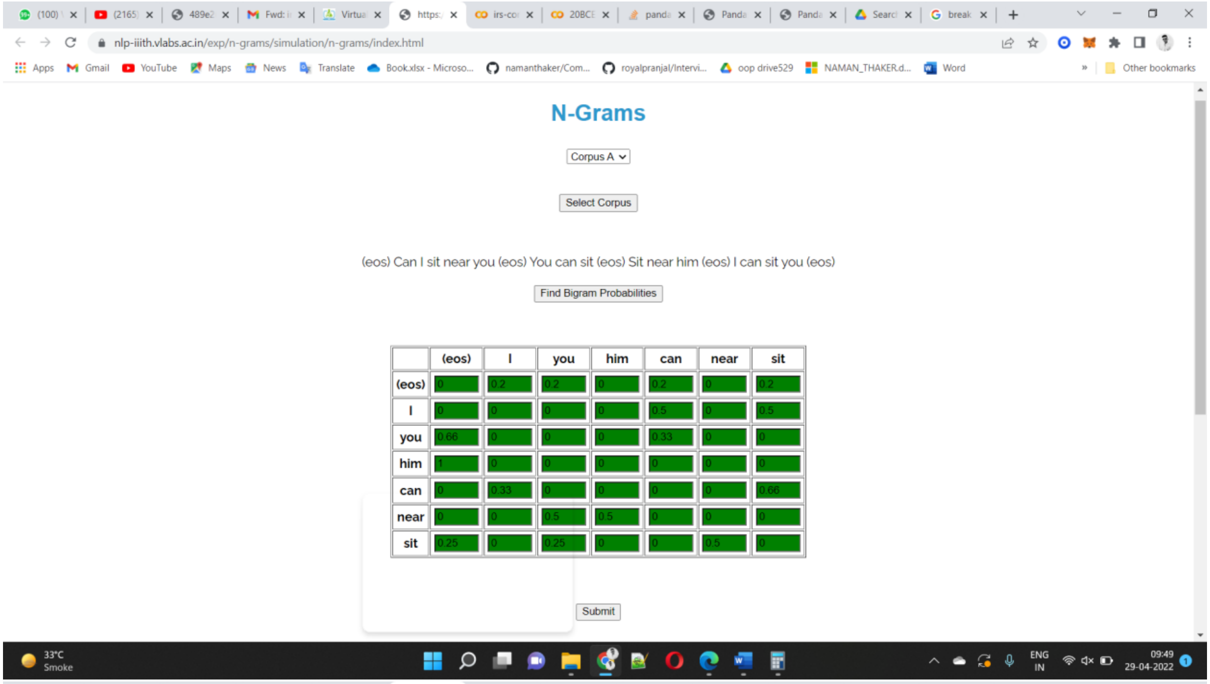
Name and Roll Number- Naman Thaker (20BCE529)

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.

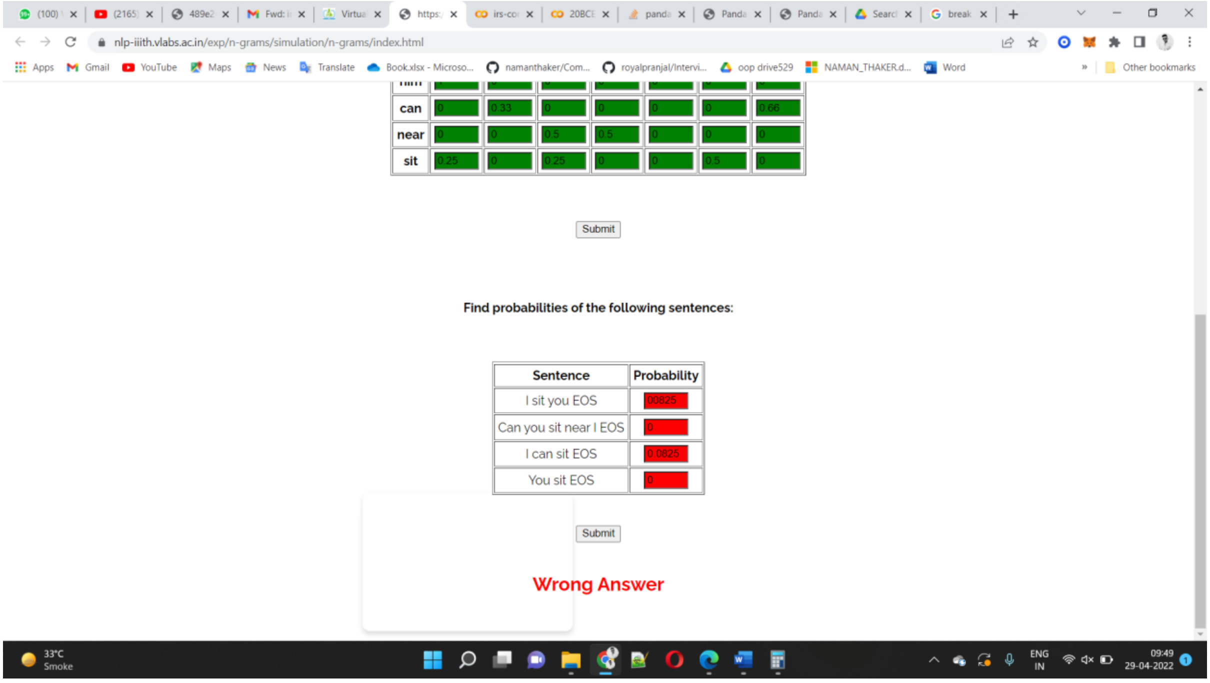
Corpus 1


nlp-iiith.vlabs.ac.in/exp/n-grams/simulation/n-grams/index.html
Corpus A
Select Corpus
Find Bigram Probabilities
Submit
33°C
Smoke
09:49
29-04-2022

N-Grams

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

	(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


nlp-iiith.vlabs.ac.in/exp/n-grams/simulation/n-grams/index.html
Submit
Submit
33°C
Smoke
09:49
29-04-2022

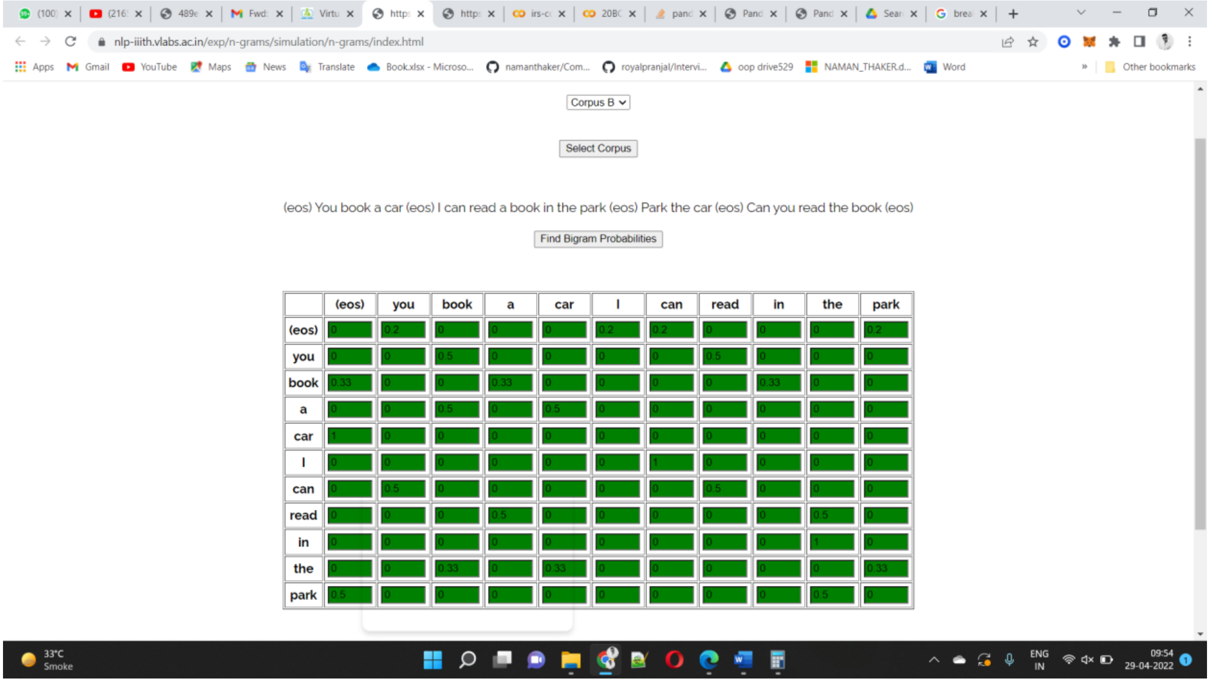
can	0	0.33	0	0	0	0	0.66
near	0	0	0.5	0.5	0	0	0
sit	0.25	0	0.25	0	0	0.5	0

Find probabilities of the following sentences:

Sentence	Probability
I sit you EOS	00825
Can you sit near I EOS	0
I can sit EOS	0.0825
You sit EOS	0

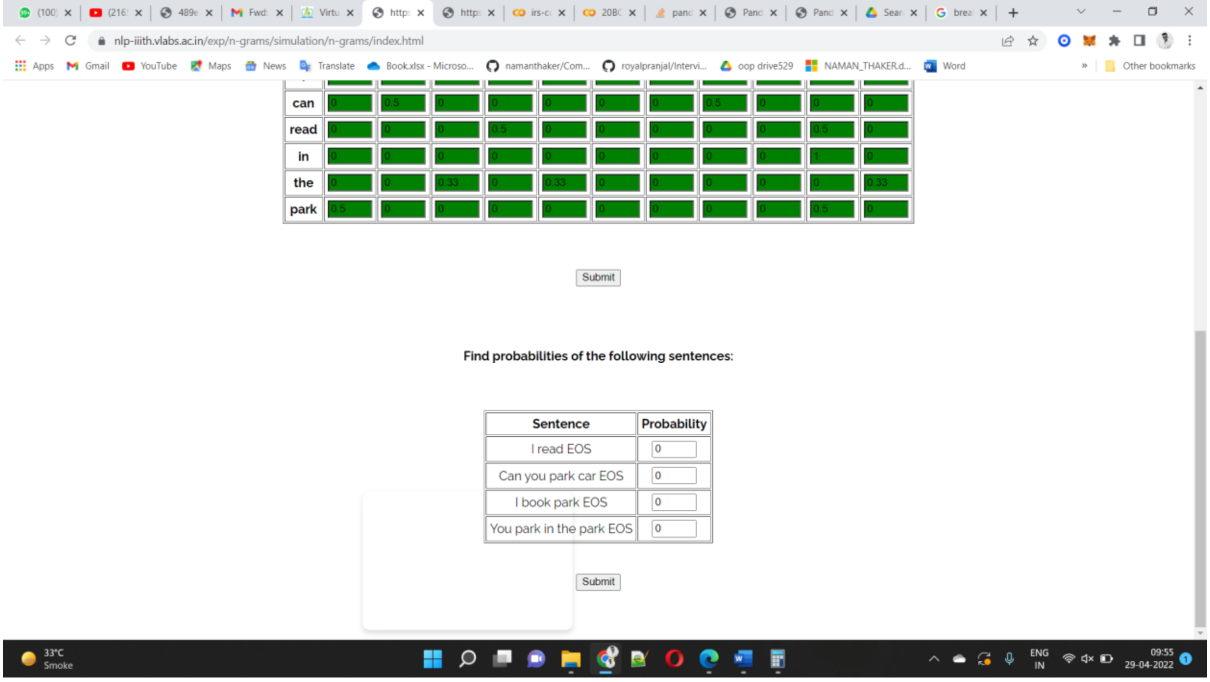
Wrong Answer

Corpus 2


nlp-iiith.vlabs.ac.in/exp/n-grams/simulation/n-grams/index.html
Corpus B
Select Corpus
Find Bigram Probabilities

(eos) You book a car (eos) I can read a book in the park (eos) Park the car (eos) Can you read the book (eos)

	(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


nlp-iiith.vlabs.ac.in/exp/n-grams/simulation/n-grams/index.html
Submit
Submit

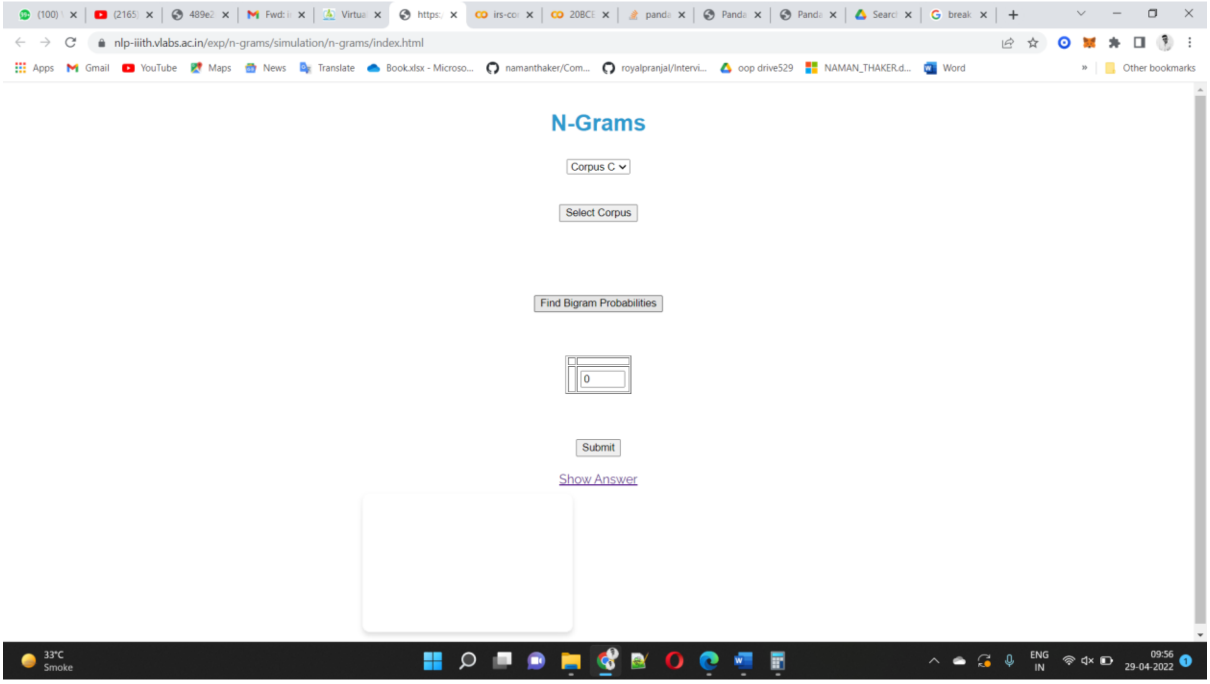
can	0	0.5	0	0	0	0	0	0.5	0	0	0
read	0	0	0	0.5	0	0	0	0	0	0.5	0
in	0	0	0	0	0	0	0	0	0	1	0
the	0	0	0.33	0	0.33	0	0	0	0	0	0.33
park	0.5	0	0	0	0	0	0	0	0	0.5	0

Find probabilities of the following sentences:

Sentence	Probability
I read EOS	0
Can you park car EOS	0
I book park EOS	0
You park in the park EOS	0

Corpus 3

No corpus given


nlp-iiith.vlabs.ac.in/exp/n-grams/simulation/n-grams/index.html
N-Grams
Corpus C
Select Corpus
Find Bigram Probabilities
0
Submit
Show Answer
33°C
Smoke
ENG
IN
09:56
29-04-2022

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

2. N Gram Smoothing

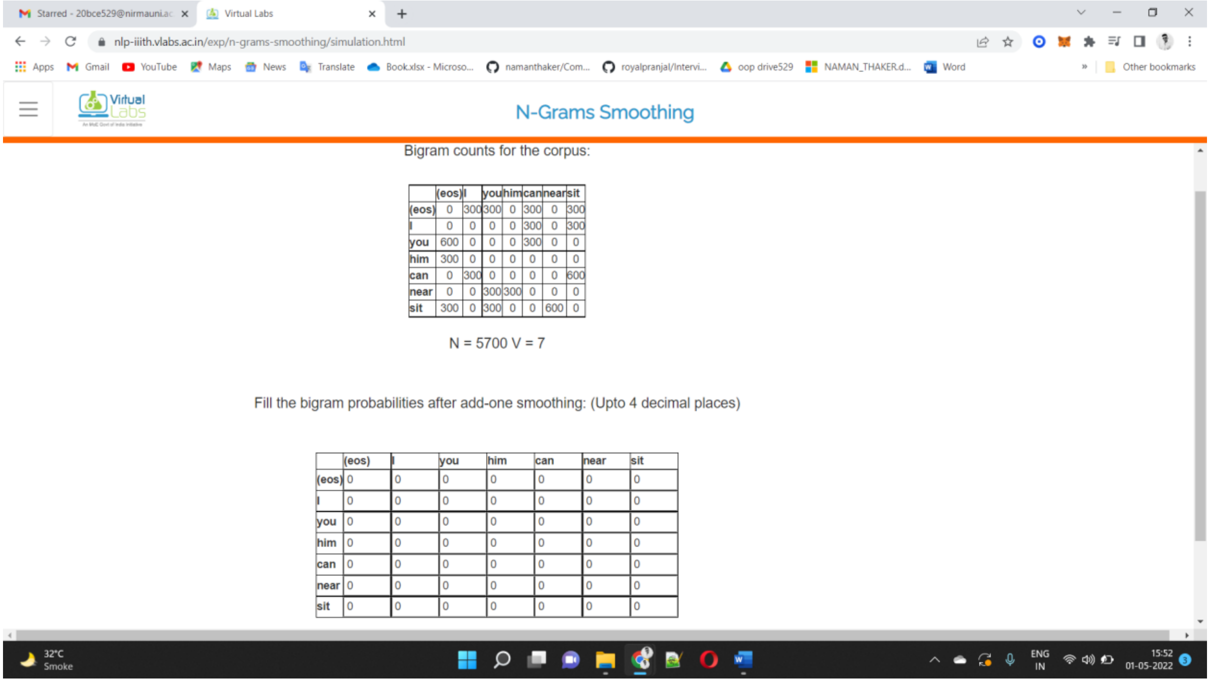
AIM: Standard N-gram models have one key flaw: they must be trained from some corpus, and because every training corpus is limited, some perfectly good N-grams are guaranteed to be missed. The bigram matrix for any given training corpus is sparse, as can be shown. There are a lot of scenarios with zero 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 :



N-Grams Smoothing

Bigram counts for the corpus:

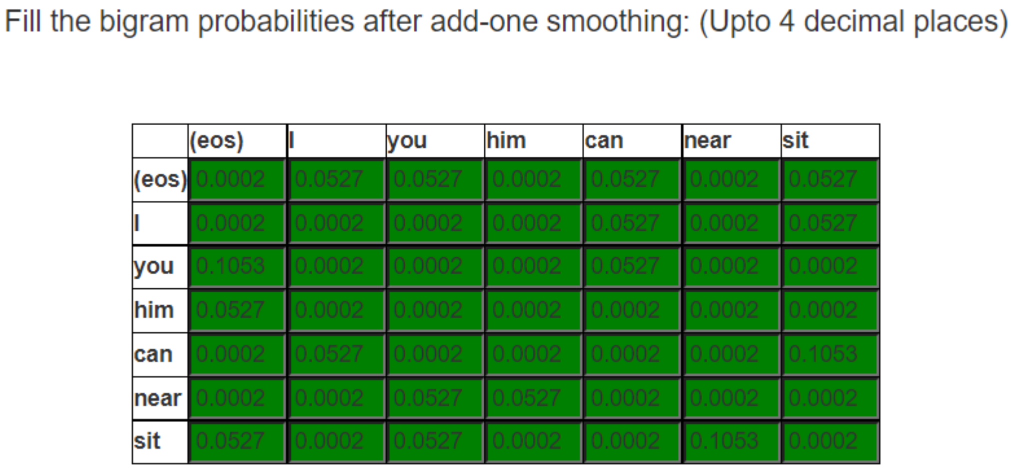
	(eos)	I	you	him	can	near	sit
(eos)	0	300	300	0	300	0	300
I	0	0	0	0	300	0	300
you	600	0	0	0	300	0	0
him	300	0	0	0	0	0	0
can	0	300	0	0	0	0	600
near	0	0	300	300	0	0	0
sit	300	0	300	0	0	600	0

N = 5700 V = 7

Fill the bigram probabilities after add-one smoothing: (Upto 4 decimal places)

	(eos)	I	you	him	can	near	sit
(eos)	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0
you	0	0	0	0	0	0	0
him	0	0	0	0	0	0	0
can	0	0	0	0	0	0	0
near	0	0	0	0	0	0	0
sit	0	0	0	0	0	0	0

Output :



Fill the bigram probabilities after add-one smoothing: (Upto 4 decimal places)

	(eos)	I	you	him	can	near	sit
(eos)	0.0002	0.0527	0.0527	0.0002	0.0527	0.0002	0.0527
I	0.0002	0.0002	0.0002	0.0002	0.0527	0.0002	0.0527
you	0.1053	0.0002	0.0002	0.0002	0.0527	0.0002	0.0002
him	0.0527	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
can	0.0002	0.0527	0.0002	0.0002	0.0002	0.0002	0.1053
near	0.0002	0.0002	0.0527	0.0527	0.0002	0.0002	0.0002
sit	0.0527	0.0002	0.0527	0.0002	0.0002	0.1053	0.0002