

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):


HOMEPARTNERSCONTACT

Computer Science and Engineering

[Introduction](#)
[Objective](#)
[List of experiments](#)
[Target Audience](#)
[Course Alignment](#)
[Feedback](#)

Natural Language Processing

Natural Language is the language written or spoken by humans in their daily life. Natural Language Processing is an interdisciplinary field dealing with human-computer interaction and computer aided processing of human language. It combines major concepts from computer science, artificial intelligence, and linguistics.


HOMEPARTNERSCONTACT

Computer Science and Engineering

[Introduction](#)
[Objective](#)
[List of experiments](#)
[Target Audience](#)
[Course Alignment](#)
[Feedback](#)

Natural Language Processing

1. Word Analysis
2. Word Generation
3. Morphology
4. N-Grams
5. N-Grams Smoothing
6. POS Tagging: Hidden Markov Model
7. POS Tagging: Viterbi Decoding
8. Building POS Tagger
9. Chunking
10. Building Chunker


HOME
PARTNERS
CONTACT

Computer Science and Engineering > Natural Language Processing > Experiments

Aim

Theory

Objective

Procedure

Simulation

Assignment

References


Feedback

N-Grams

Probability of a sentence can be calculated by the probability of sequence of words occurring in it. We can use Markov assumption, that the probability of a word in a sentence depends on the probability of the word occurring just before it. Such a model is called first order Markov model or the bigram model.

$$P(W_n | W_{n-1}) = P(W_{n-1}, W_n) / P(W_{n-1})$$

Here, W_n refers to the word token corresponding to the n th word in a sequence.


HOME
PARTNERS
CONTACT

Computer Science and Engineering > Natural Language Processing > Experiments

Aim

Theory

Objective

Procedure

Simulation

Assignment

References

Feedback

N-Grams

STEP 1: Select a corpus and click on **Generate bigram table**

STEP 2: Fill up the table that is generated and hit **Submit**

STEP 3: If incorrect (red), see the correct answer by clicking on show answer or repeat Step 2.

STEP 4: If correct (green), click on take a quiz and fill the correct answer

CORPUS A:

N-Grams

Corpus A ▼

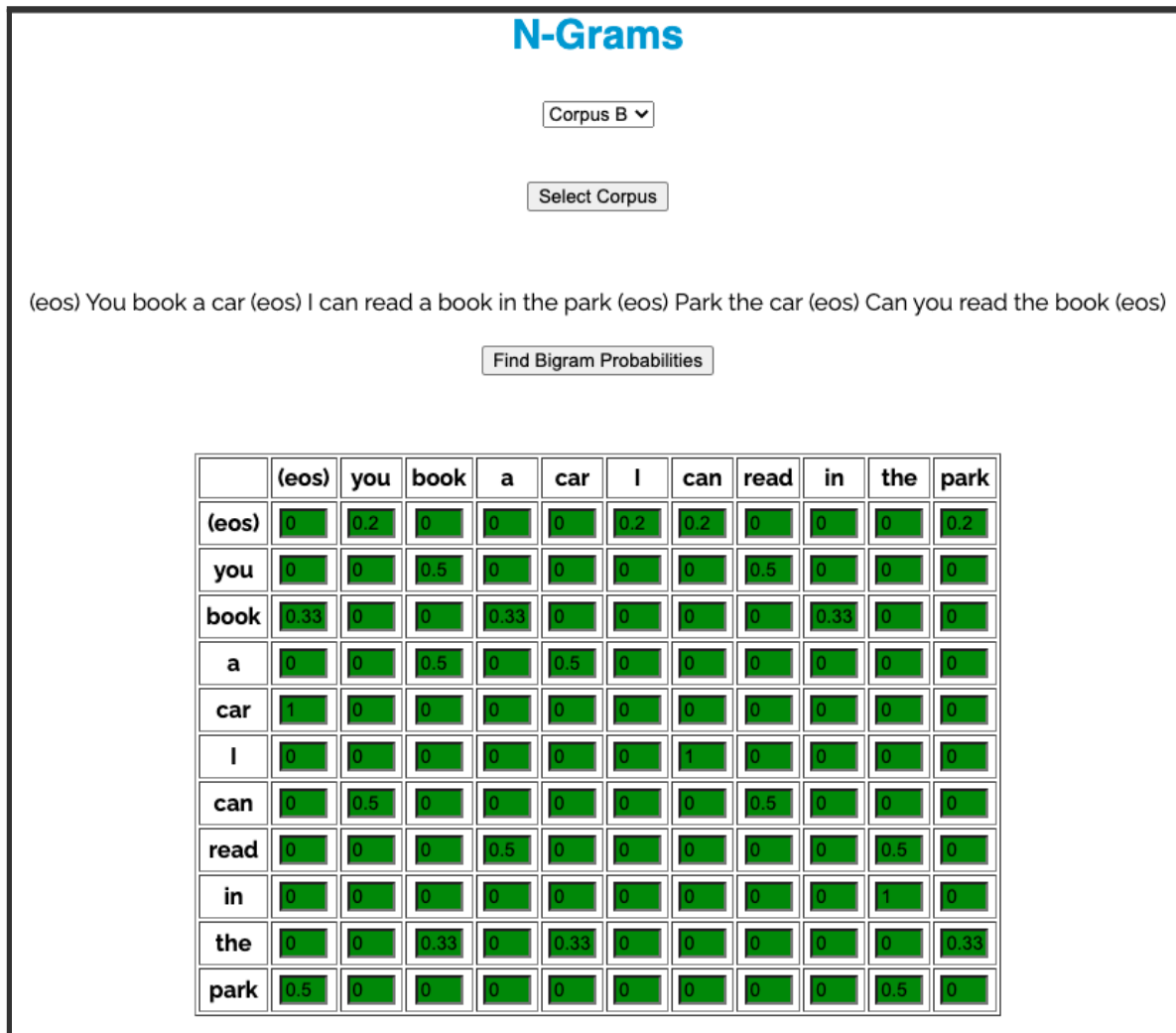
Select Corpus

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

Find Bigram Probabilities

	(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:



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 corpora 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