Terna Engineering College
Computer Engineering Department
Program: Sem VIII

Course: Natural Language Processing

Experiment No. 4

A.1 Aim: Perform and analyse smoothing operations for n-gram models using the 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: 14/02/2022	**Date of Submission:** 14/02/2022
Grade:	

B.1 Virtual Lab (Input & Output):

Computer Science and Engineering > Natural Language Processing > Experiments

Aim
Theory
Objective
Procedure
Simulation
Assignment
References
Feedback

N-Grams Smoothing

One major problem with standard N-gram models is that they must be trained from some corpus, and because any particular training corpus is finite, some perfectly acceptable N-grams are bound to be missing from it. We can see that bigram matrix for any given training corpus is sparse. There are large number of cases with zero probabilty bigrams and that should really have some non-zero probability. This method tend to underestimate the probability of strings that happen not to have occurred nearby in their training corpus.

There are some techniques that can be used for assigning a non-zero probabilty to these 'zero probability bigrams'. This task of reevaluating some of the zero-probability and low-probabilty N-grams, and assigning them non-zero values, is called smoothing.

	eos	I	booked	a	flight	took
eos	0	300	0	0	0	300
I	0	0	300	0	0	0
booked	0	0	0	300	0	0
a	0	0	0	0	600	0
flight	600	0	0	0	0	0
took	0	0	0	300	0	0

Valid bigrams absent in the training corpus:
How could I eos I have a booked room eos I took a flight eos

Computer Science and Engineering > Natural Language Processing > Experiments

Aim
Theory
Objective
Procedure
Simulation
Assignment
References
Feedback

N-Grams Smoothing

STEP 1: Select a corpus

STEP 2: Apply add one smoothing and calculate bigram probabilities using the given bigram counts,N and V. Fill the table and hit `Submit`

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

N-Grams Smoothing

Corpus A

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)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
I	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
you	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
him	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
can	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
near	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
sit	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Submit

Right Answer

Corpus A

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)	0.0002	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

Submit

	(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

Submit

Right Answer

B.2 Observations and learning:

- The bigram model uses just the conditional probability of one preceding word to estimate the likelihood of a word given all previous words. In other words, you use probability to approximate it: P(the | that)

B.3 Conclusion:

As a result, we've figured out how to apply add-one smoothing to a sparse bigram table and put it into practice.

B.4 Question of Curiosity

Q1. Add-one smoothing works horribly in practice because of giving too much probability mass to unseen n-grams. Prove using an example.

ANS:

Add-one smoothing

p(wi |wi−1) = 1 + c(wi−1wi)/ P wi [1 + c(wi−1wi)] = 1 + c(wi−1wi) / |V | + P wi c(wi−1wi)

- Originally due to Laplace.
- Typically, we assume V = {w : c(w) > 0} ∪ {UNK}
- Add-one smoothing is generally a horrible choice.

JOHN READ MOBY DICKENS
MARY READ A DIFFERENT BOOK
SHE READ A BOOK BY CHER

p(JOHN READ A BOOK) = 1+1 11+3 1+1 11+1 1+2 11+3 1+1 11+2 1+1 11+2 ≈ 0.0001 p(CHER READ A BOOK) = 1+0 11+3 1+0 11+1 1+2 11+3 1+1 11+2 1+1 11+2 ≈ 0.00003

Q2. In Add-δ smoothing, we add a small value 'δ' to the counts instead of one. Apply Add-δ smoothing to the below bigram count table where δ=0.02.

ANS:

	(eos)	John	read	Fountainhead	Mary	a	different	book	She	by	Dickens
(eos)	0	300	0	0	300	0	0	0	300	0	0
John	0	0	300	0	0	0	0	0	0	0	0
read	0	0	0	300	0	600	0	0	0	0	0
Fountainhead	300	0	0	0	0	0	0	0	0	0	0
Mary	0	0	300	0	0	0	0	0	0	0	0
a	0	0	0	0	0	0	300	300	0	0	0
different	0	0	0	0	0	0	0	300	0	0	0
book	300	0	0	0	0	0	0	0	0	300	0
She	0	0	0	300	0	0	0	0	0	0	0
by	0	0	0	0	0	0	0	0	0	0	300
Dickens	300	0	0	0	0	0	0	0	0	0	0

N = 5100 V = 11

Q3. Given S = Dickens read a book, find P(S)

(a) using unsmoothed probability

(b) applying Add-One smoothing.

(c) applying Add-δ smoothing

ANS:

	(eos)	John	read	Fountainhead	Marry	a	different	book	She	by	Dickens
(eos)	0.0003	0.0527	0.0527	0.0003	0.0003	0.0003	0.0003	0.0003	0.0527	0.0003	0.0527
John	0.0003	0.0003	0.0003	0.0003	0.0527	0.0003	0.0003	0.0003	0.0003	0.0003	0.0527
read	0.1053	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0527	0.0003	0.0003
Fountainhead	0.0003	0.0003	0.0003	0.0527	0.0527	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
Marry	0.0003	0.0003	0.0003	0.0527	0.0003	0.0527	0.0003	0.0003	0.0003	0.0003	0.1053
a	0.0003	0.0003	0.0003	0.0527	0.0003	0.0003	0.0527	0.0527	0.0003	0.0003	0.0003
different	0.0527	0.0003	0.0527	0.0003	0.0003	0.1053	0.0003	0.0003	0.0003	0.0003	0.0003
book	0.0527	0.0003	0.0527	0.0003	0.0003	0.1053	0.0003	0.0003	0.0003	0.0003	0.0003
She	0.0527	0.0003	0.0527	0.0003	0.0003	0.1053	0.0003	0.0003	0.0003	0.0003	0.0003
by	0.0527	0.0003	0.0527	0.0003	0.0003	0.1053	0.0003	0.0003	0.0003	0.0003	0.0003
Dickens	0.0527	0.0003	0.0527	0.0003	0.0003	0.1053	0.0003	0.0003	0.0003	0.0003	0.0003