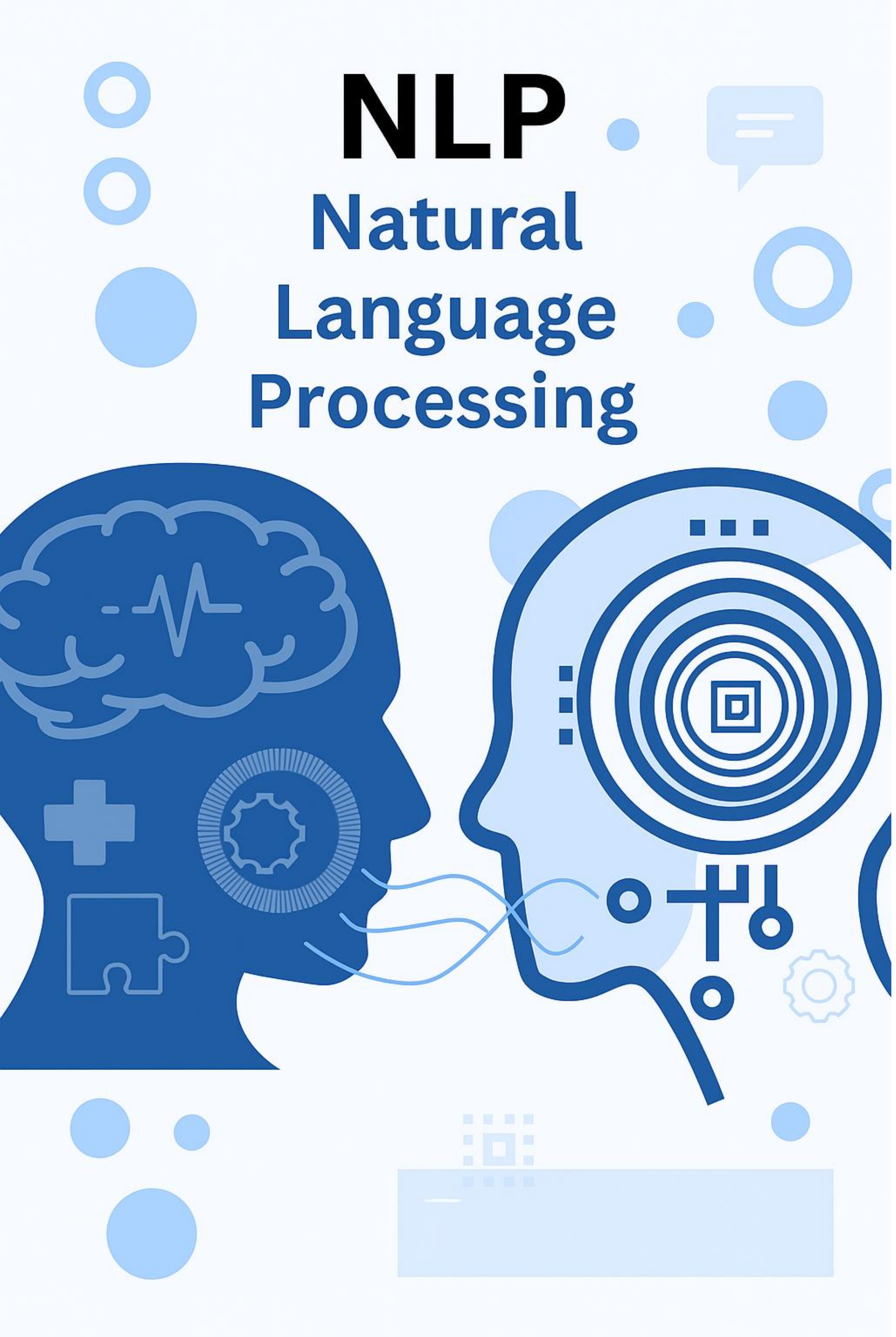

NATURAL LANGUAGE PROCESSING (NLP)

PMDS606L

MODULE 3

LECTURE 3

Dr. Kamanasish Bhattacharjee

Assistant Professor

Dept. of Analytics, SCOPE, VIT

TRAINING CORPUS

I love NLP

I love AI

AI is powerful

BIGRAMS

Bigram	Count
(I, love)	2
(love, NLP)	1
(love, AI)	1
(AI, is)	1
(is, powerful)	1

SMOOTHING

The standard way to deal with putative “zero probability n-grams” that should really have some non-zero probability is called smoothing or discounting. Smoothing algorithms shave off a bit of probability mass from some more frequent events and give it to unseen events.

LAPLACE (ADD-ONE) SMOOTHING

$$P(w_i) = \frac{c_i}{N}$$

Unigram Smoothing –

$$P_{\text{Laplace}}(w_i) = \frac{c_i + 1}{N + V}$$

TRAINING CORPUS

I love programming

Programmin is fun

Fun is subjective

UNIGRAM COUNTS

Unigram	Frequency	Probability
I	1	$1/9 \approx 0.111$
love	1	$1/9 \approx 0.111$
programming	2	$2/9 \approx 0.222$
is	2	$2/9 \approx 0.222$
fun	2	$2/9 \approx 0.222$
subjective	1	$1/9 \approx 0.111$
Happiness	0	$0/9 = 0$
Coding	0	$0/9 = 0$
Algorithm	0	$0/9 = 0$

UNIGRAM COUNTS WITH LAPLACE SMOOTHING

Unigram	Frequency	Probability P(Wi) = (Count(Wi) + 1) / (Total_Words + V)		
I	1	(1 + 1) / (9 + 9)	= 2 / 18	≈ 0.111
love	1	(1 + 1) / (9 + 9)	= 2 / 18	≈ 0.111
programming	2	(2 + 1) / (9 + 9)	= 3 / 18	= 0.167
is	2	(2 + 1) / (9 + 9)	= 3 / 18	= 0.167
fun	2	(2 + 1) / (9 + 9)	= 3 / 18	= 0.167
subjective	1	(1 + 1) / (9 + 9)	= 2 / 18	≈ 0.111
Happiness	0	(0 + 1) / (9 + 9)	= 1 / 18	≈ 0.056
Coding	0	(0 + 1) / (9 + 9)	= 1 / 18	≈ 0.056
Algorithm	0	(0 + 1) / (9 + 9)	= 1 / 18	≈ 0.056

LAPLACE (ADD-ONE) SMOOTHING

$$P_{\text{MLE}}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

Bigram Smoothing –

$$P_{\text{Laplace}}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{\sum_w (C(w_{n-1}w)+1)} = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V}$$

BIGRAM COUNTS

Bigram	Frequency	Probability P(w1, w2) = Count(w1, w2) / Count(w1)		
(“I”, “love”)	1	Count(“I”, “love”) / Count(“I”)	= 1/1	= 1
(“love”, “programming”)	1	Count(“love”, “programming”) / Count(“love”)	= 1/1	= 1
(“programming”, “is”)	1	Count("programming", "is") / Count("programming")	= 1/2	= 0.5
(“is”, “fun”)	1	Count("is", "fun") / Count("is")	= 1/2	= 0.5
(“fun”, “is”)	1	Count("fun", "is") / Count("fun")	= 1/2	= 0.5
(“is”, “subjective”)	1	Count(“is”, “subjective”) / Count(“is”)	= 1/2	= 0.5

UNIGRAM AND BIGRAM COUNTS

Unigram	Frequency
Chicago	4
is	8
cold	6
hot	0

Bigram	Frequency
Chicago is	2
is cold	4
is hot	0
…	0

UNIGRAM AND BIGRAM PROBABILITY MATRIX

Unigram	Probability
Chicago	$\frac{4}{18} = 0.22$
is	$\frac{8}{18} = 0.44$
cold	$\frac{6}{18} = 0.33$
hot	$\frac{0}{18} = 0.00$

Bigram	Probability
Chicago is	$\frac{2}{4} = 0.50$
is cold	$\frac{4}{8} = 0.50$
is hot	$\frac{0}{8} = 0.00$

BEFORE AND AFTER LAPLACE SMOOTHING

Bigram	Probability
Chicago is	$\frac{2}{4} = 0.50$
is cold	$\frac{4}{8} = 0.50$
is hot	$\frac{0}{8} = 0.00$

Bigram	Probability
Chicago is	$\frac{3}{8} = 0.38$
is cold	$\frac{5}{12} = 0.42$
is hot	$\frac{1}{12} = 0.08$

BERKELEY RESTAURANT PROJECT

A dialogue system from the last century that answered questions about a database of restaurants in Berkeley, California.

9332 sentences

1446 words

can you tell me about any good cantonese restaurants close by
tell me about chez panisse
i'm looking for a good place to eat breakfast
when is caffe venezia open during the day

BIGRAM AND UNIGRAM COUNTS

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

BIGRAM COUNTS WITH LAPLACE SMOOTHING

	i	want	to	eat	chinese	food	lunch	spend
i	6	828	1	10	1	1	1	3
want	3	1	609	2	7	7	6	2
to	3	1	5	687	3	1	7	212
eat	1	1	3	1	17	3	43	1
chinese	2	1	1	1	1	83	2	1
food	16	1	16	1	2	5	1	1
lunch	3	1	1	1	1	2	1	1
spend	2	1	2	1	1	1	1	1

BIGRAM PROABILITY MATRIX NORMALIZED BY UNIGRAM COUNTS

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

BIGRAM PROABILITY MATRIX NORMALIZED BY UNIGRAM COUNTS WITH LAPLACE SMOOTHING

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

BIGRAM COUNTS

	(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

BIGRAM COUNTS WITH LAPLACE SMOOTHING

	(eos)	I	you	him	can	near	sit
(eos)	1	301	301	1	301	1	301
I	1	1	1	1	301	1	301
you	601	1	1	1	301	1	1
him	301	1	1	1	1	1	1
can	1	301	1	1	1	1	601
near	1	1	301	301	1	1	1
sit	301	1	301	1	1	601	1

BIGRAM COUNTS WITH LAPLACE SMOOTHING

	(eos)	I	you	him	can	near	sit
(eos)	0.0008	0.2479	0.2479	0.0008	0.2479	0.0008	0.2479
I	0.0016	0.0016	0.0016	0.0016	0.4902	0.0016	0.4902
you	0.6575	0.0011	0.0011	0.0011	0.3293	0.0011	0.0011
him	0.9586	0.0032	0.0032	0.0032	0.0032	0.0032	0.0032
can	0.0011	0.3293	0.0011	0.0011	0.0011	0.0011	0.6575
near	0.0016	0.0016	0.4902	0.4902	0.0016	0.0016	0.0016
sit	0.2479	0.0008	0.2479	0.0008	0.0008	0.4951	0.0008

ADD-k SMOOTHING

$$P_{\text{MLE}}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

Add-k Smoothing –

$$P^{*}_{\text{Add-k}}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+k}{C(w_{n-1})+kV}$$

ADD-k SMOOTHING

“Apple”: 5

“Banana”: 3

“Cherry”: 2

“Dates”: 0 (Not present in the corpus)

The vocabulary size is 4 (including “Apple,” “Banana,” “Cherry,” and “Dates”).

WITHOUT SMOOTHING

$$P_{\text{no-smoothing}}(\text{"Apple"}) = \frac{\text{Count("Apple")}}{\text{Total Count}} = \frac{5}{10} = 0.5$$

$$P_{\text{no-smoothing}}(\text{"Banana"}) = \frac{\text{Count("Banana")}}{\text{Total Count}} = \frac{3}{10} = 0.3$$

$$P_{\text{no-smoothing}}(\text{"Cherry"}) = \frac{\text{Count("Cherry")}}{\text{Total Count}} = \frac{2}{10} = 0.2$$

$$P_{\text{no-smoothing}}(\text{"Dates"}) = \frac{\text{Count("Dates")}}{\text{Total Count}} = \frac{0}{10} = 0$$

LAPLACE SMOOTHING

$$P_{\text{laplace}}(\text{"Apple"}) = \frac{\text{Count}(\text{"Apple"})+1}{\text{Total Count}+V} = \frac{5+1}{10+4} = \frac{6}{14} \approx 0.4286$$

$$P_{\text{laplace}}(\text{"Banana"}) = \frac{\text{Count}(\text{"Banana"})+1}{\text{Total Count}+V} = \frac{3+1}{10+4} = \frac{4}{14} \approx 0.2857$$

$$P_{\text{laplace}}(\text{"Cherry"}) = \frac{\text{Count}(\text{"Cherry"})+1}{\text{Total Count}+V} = \frac{2+1}{10+4} = \frac{3}{14} \approx 0.2143$$

$$P_{\text{laplace}}(\text{"Dates"}) = \frac{\text{Count}(\text{"Dates"})+1}{\text{Total Count}+V} = \frac{0+1}{10+4} = \frac{1}{14} \approx 0.0714$$

Use Add-0.5 Smoothing