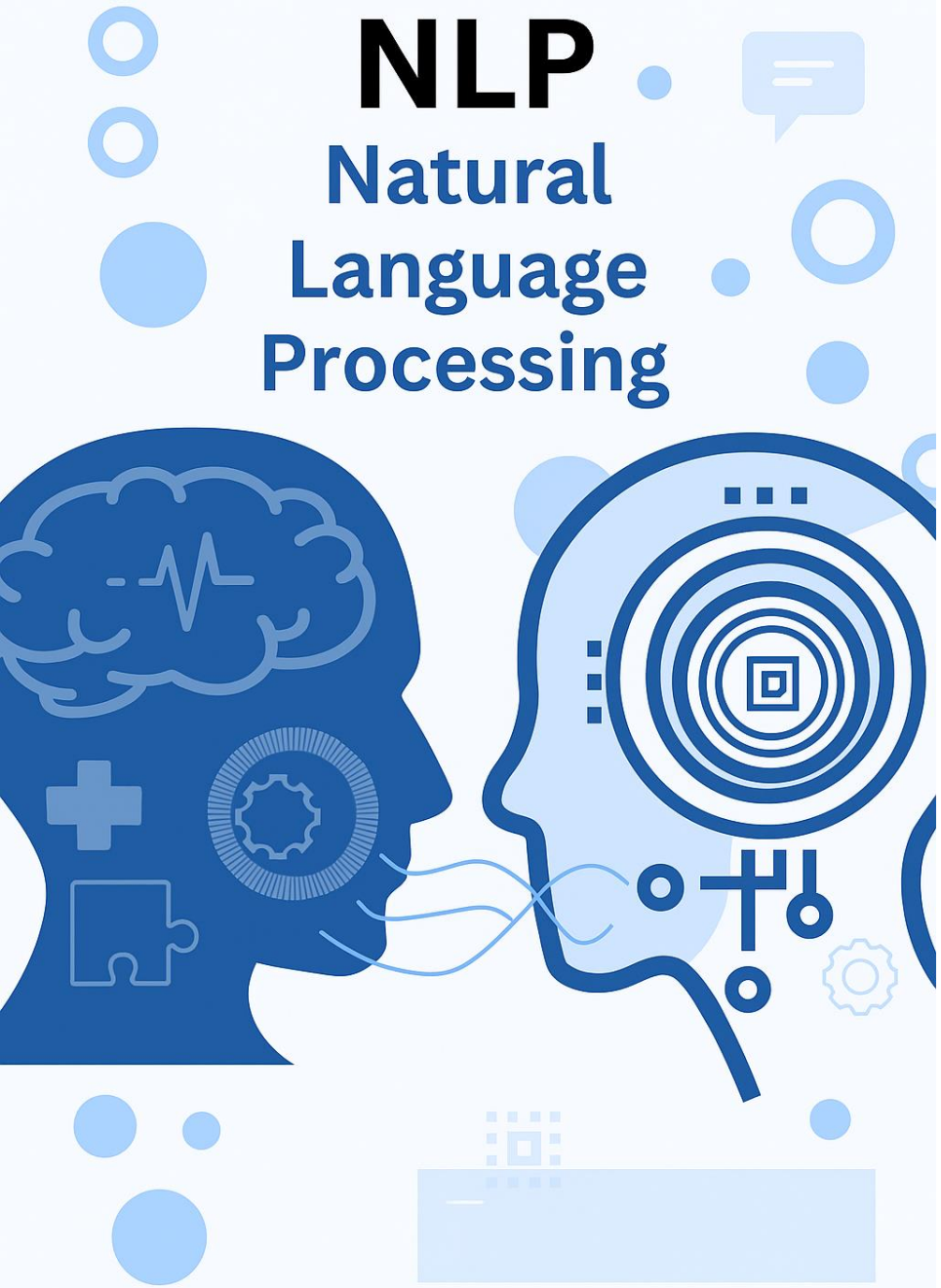

NATURAL LANGUAGE PROCESSING (NLP)

PMDS606L

MODULE 3

LECTURE 2

Dr. Kamanasish Bhattacharjee

Assistant Professor

Dept. of Analytics, SCOPE, VIT

UNIGRAM

The dog barks

The cat sleeps

The dog runs

The cat jumps

P(dog)?

TRIGRAM

The girl bought a chocolate

The boy ate the chocolate

The girl bought a toy

The girl played with the toy

P(The girl bought)?

P(The girl played)?

N-GRAM

I am Henry

I like college

Do Henry like college

Henry I am

Do I like Henry

Do I like college

I do like Henry

Bigram
Do _____ ?
Henry _____ ?

Trigram
I like _____ ?
Do I _____ ?

4-gram
I like college ___ ?
Do I like _____ ?

P (I like college)?
P (Do I like Henry)?

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 UNIRAM PROABILITY MATRIX

	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 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

BERKELEY RESTAURANT PROJECT

Here are a few other useful probabilities:

$$P(\texttt{i}\,|\,\texttt{<s>}) = 0.25 \qquad P(\texttt{english}|\texttt{want}) = 0.0011$$
$$P(\texttt{food}|\texttt{english}) = 0.5 \qquad P(\texttt{</s>}|\texttt{food}) = 0.68$$

Now we can compute the probability of sentences like *I want English food* or *I want Chinese food* by simply multiplying the appropriate bigram probabilities together, as follows:

$$\begin{aligned} P(\texttt{<s> i want english food </s>}) &= P(\texttt{i}|\texttt{<s>})P(\texttt{want}|\texttt{i})P(\texttt{english}|\texttt{want}) \\ &\quad P(\texttt{food}|\texttt{english})P(\texttt{</s>}|\texttt{food}) \\ &= 0.25 \times 0.33 \times 0.0011 \times 0.5 \times 0.68 \\ &= 0.000031 \end{aligned}$$

BIGRAM

I am from Vellore

I am a teacher

students are good and are from various cities

students from Vellore do engineering

BIGRAM PROABILITY MATRIX NORMALIZED BY UNIGRAM COUNTS

		w_n				
		students	are	from	Vellore	</s>
w_{n-1}	<s>	2/4	0/4	0/4	0/4	0/4
	students	0/2	1/2	1/2	0/2	0/2
	are	0/2	0/2	1/2	0/2	0/2
	from	0/3	0/3	0/3	2/3	0/3
	Vellore	0/2	0/2	0/2	0/2	1/2