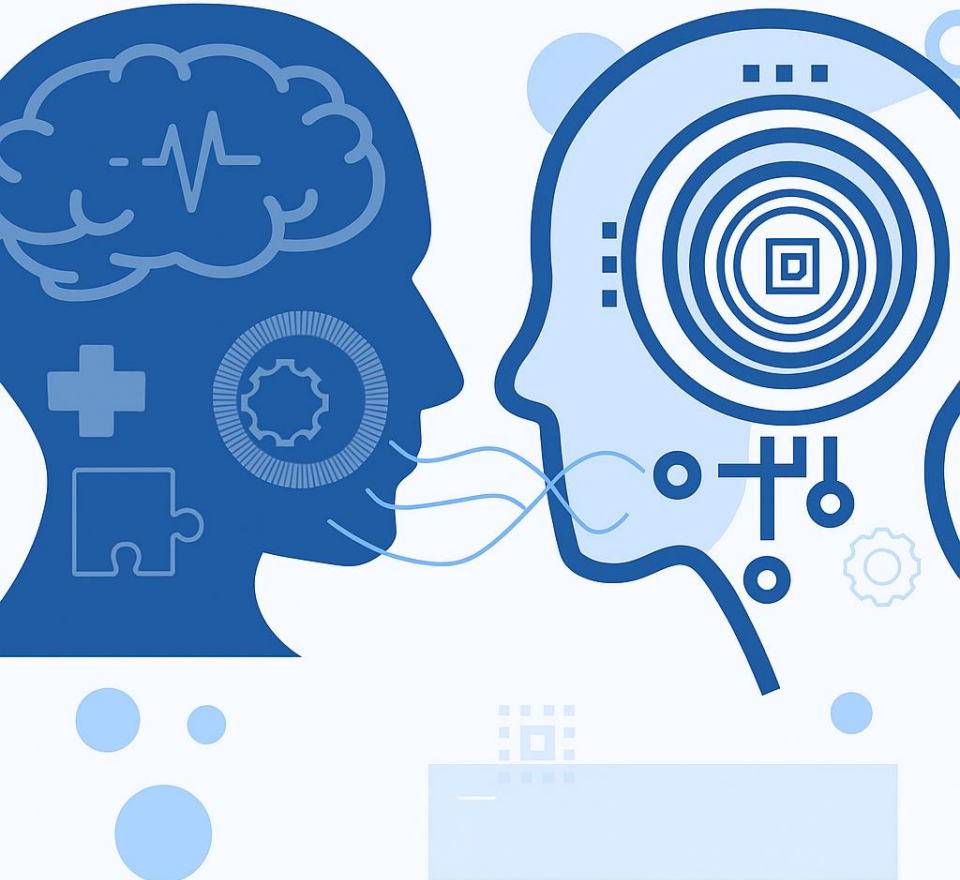


# NLP

Natural  
Language  
Processing



# NATURAL LANGUAGE PROCESSING (NLP)

## PMDS606L

MODULE 3

LECTURE 1

***Dr. Kamanasish Bhattacharjee***

*Assistant Professor*

*Dept. of Analytics, SCOPE, VIT*



***Please turn your homework .....***

You would not believe the  
day I had

I devoured a baby in a cab

You WHAT!!!

Ha! Oh god. Delivered!!!  
This phone I swear

OMG wow boy or girl

Gorilla

GIRL

# LANGUAGE MODEL

We formalize this idea of word prediction with probabilistic models called N-gram models, which predict the next word from the previous  $N - 1$  words. Such statistical models of word sequences are called **Language Models (LMs)**.

A **language model** assigns a **probability** to a sequence of words and helps determine **what comes next** in a sentence.

Suppose we have:

***"The cat sat on the"***

A language model might predict:

"mat" (high probability)

"tree" (lower probability)

"rocket" (very low probability)

# LANGUAGE MODEL

The following sequence has a non-zero probability of appearing in a text:

***.....all of a sudden I notice three guys standing on the sidewalk...***

while this same set of words in a different order has a very low probability:

***on guys all I of notice sidewalk three a sudden standing the***

# N-Gram

N=1:

This is a sentence

Uni-grams

this,  
is,  
a,  
sentence

N=2:

This is a sentence

Bi-grams

this is,  
is a,  
a sentence

N=3:

This is a sentence

Tri-grams

this is a,  
is a sentence

# APPLICATIONS OF LMs

- **Text Generation** (e.g., GPT-4)
- **Text Completion/Autocompletion** (e.g., Google Search, email suggestions)
- **Machine Translation** (e.g., English → Hindi)
- **Speech Recognition & Synthesis**
- **Spelling/Grammar Correction**
- **Question Answering & Chatbots**

# USING N-GRAMS TO PREDICT THE PROBABILITY OF A SENTENCE

For every sentence, we put  $\langle s \rangle$  and  $\langle /s \rangle$  at the beginning and the end respectively. This denote the start and the end of the sentence.

Corpus:

$\langle s \rangle$  I am a human  $\langle /s \rangle$

$\langle s \rangle$  I am not a stone  $\langle /s \rangle$

$\langle s \rangle$  I I live in Mumbai  $\langle /s \rangle$

Check the probability of "II am not"  
using bigram

Read as  
Prob of "am" given "I"

$$P(\text{II am not}) = P(I/\langle s \rangle) \times P(I/I) \times P(\text{am}/I) \times P(\text{not}/\text{am}) \times P(\langle /s \rangle/\text{not})$$
$$= \frac{\text{Count}(\langle s \rangle I)}{\text{count}(\langle s \rangle)} \times \frac{\text{Count}(I/I)}{\text{count}(I)} \times \frac{\text{Count}(I|\text{am})}{\text{count}(I)} \times \frac{\text{Count}(\text{am}|\text{not})}{\text{Count}(\text{am})} \times \frac{\text{Count}(\text{not}|\langle /s \rangle)}{\text{count}(\langle /s \rangle)}$$

$\Rightarrow \text{Count}(\langle s \rangle I) \Rightarrow$  In our corpus, we have to check the frequency of the combination  $\langle s \rangle I$  and that in our corpus is 3

$$\text{Count}(\langle s \rangle) = 3$$

$$\frac{3}{3} \times \frac{1}{4} \times \frac{2}{4} \times \frac{1}{2} \times \frac{0}{3} = 0$$



# USING N-GRAMS TO PREDICT THE NEXT WORD

For every sentence, we put  $\langle s \rangle$  and  $\langle /s \rangle$  at the beginning and the end respectively. This denote the start and the end of the sentence.

Consider the following training data

$\langle s \rangle$  I am Jack  $\langle /s \rangle$

$\langle s \rangle$  Jack I am  $\langle /s \rangle$

$\langle s \rangle$  Jack I like  $\langle /s \rangle$

$\langle s \rangle$  Jack I do like  $\langle /s \rangle$

$\langle s \rangle$  do I like Jack  $\langle /s \rangle$

Assume that we use a bigram language model based on the above data

What is the most probable next word predicted by model

1)  $\langle s \rangle$  Jack \_\_\_\_\_

2)  $\langle s \rangle$  Jack I do \_\_\_\_\_

3)  $\langle s \rangle$  Jack I am Jack \_\_\_\_\_

4)  $\langle s \rangle$  do I like \_\_\_\_\_

# USING N-GRAMS TO PREDICT THE NEXT WORD

$$P(I|\langle s \rangle) = \frac{\text{Count}(\langle s \rangle | I)}{\text{Count}(\langle s \rangle)} = \frac{1}{5}$$

$$P(\text{am}|I) = \frac{\text{Count}(I | \text{am})}{\text{Count}(I)} = \frac{2}{5}$$

$$P(\text{Jack}|\text{am}) = \frac{\text{Count}(\text{am} | \text{Jack})}{\text{Count}(\text{am})} = \frac{1}{2}$$

$$P(\langle s \rangle | \text{Jack}) = \frac{\text{Count}(\text{Jack} | \langle s \rangle)}{\text{Count}(\text{Jack})} = \frac{2}{5}$$

$$P(\text{Jack}|\langle s \rangle) = \frac{\text{Count}(\langle s \rangle | \text{Jack})}{\text{Count}(\langle s \rangle)} = \frac{2}{5}$$

$$P(I|\text{Jack}) = \frac{\text{Count}(\text{Jack} | I)}{\text{Count}(\text{Jack})} = \frac{3}{5}$$

$$P(\langle s \rangle | \text{am}) = \frac{\text{Count}(\text{am} | \langle s \rangle)}{\text{Count}(\text{am})} = \frac{1}{2}$$

$$P(\text{like}|I) = \frac{\text{Count}(I | \text{like})}{\text{Count}(I)} = \frac{2}{5}$$

$$P(\langle s \rangle | \text{like}) = \frac{\text{Count}(\text{like} | \langle s \rangle)}{\text{Count}(\text{like})} = \frac{2}{3}$$

$$P(\text{do}|I) = \frac{\text{Count}(I | \text{do})}{\text{Count}(I)} = \frac{1}{5}$$

$$P(\text{like}|\text{do}) = \frac{\text{Count}(\text{do} | \text{like})}{\text{Count}(\text{do})} = \frac{1}{2}$$

$$P(\text{do}|\langle s \rangle) = \frac{\text{Count}(\langle s \rangle | \text{do})}{\text{Count}(\langle s \rangle)} = \frac{1}{5}$$

$$P(I|\text{do}) = \frac{\text{Count}(\text{do} | I)}{\text{Count}(\text{do})} = \frac{1}{2}$$

$$P(\text{Jack}|\text{like}) = \frac{\text{Count}(\text{like} | \text{Jack})}{\text{Count}(\text{like})} = \frac{1}{3}$$

# USING N-GRAMS TO PREDICT THE NEXT WORD

1) Jack —

$\Rightarrow P(\text{something}|\text{Jack}) =$  In our calculated probabilities we got 2 probabilities

1)  $P(</s>|\text{Jack}) = \frac{2}{5}$

2)  $P(I|\text{Jack}) = \frac{3}{5}$

} Since  $\frac{3}{5} > \frac{2}{5}$ , I is the next word

2) Jack I do —

$P(\text{something}|\text{do}) \rightarrow \begin{cases} P(I|\text{do}) = 1/2 \\ P(\text{like}|\text{do}) = 1/2 \end{cases}$  } The answer is both I and like