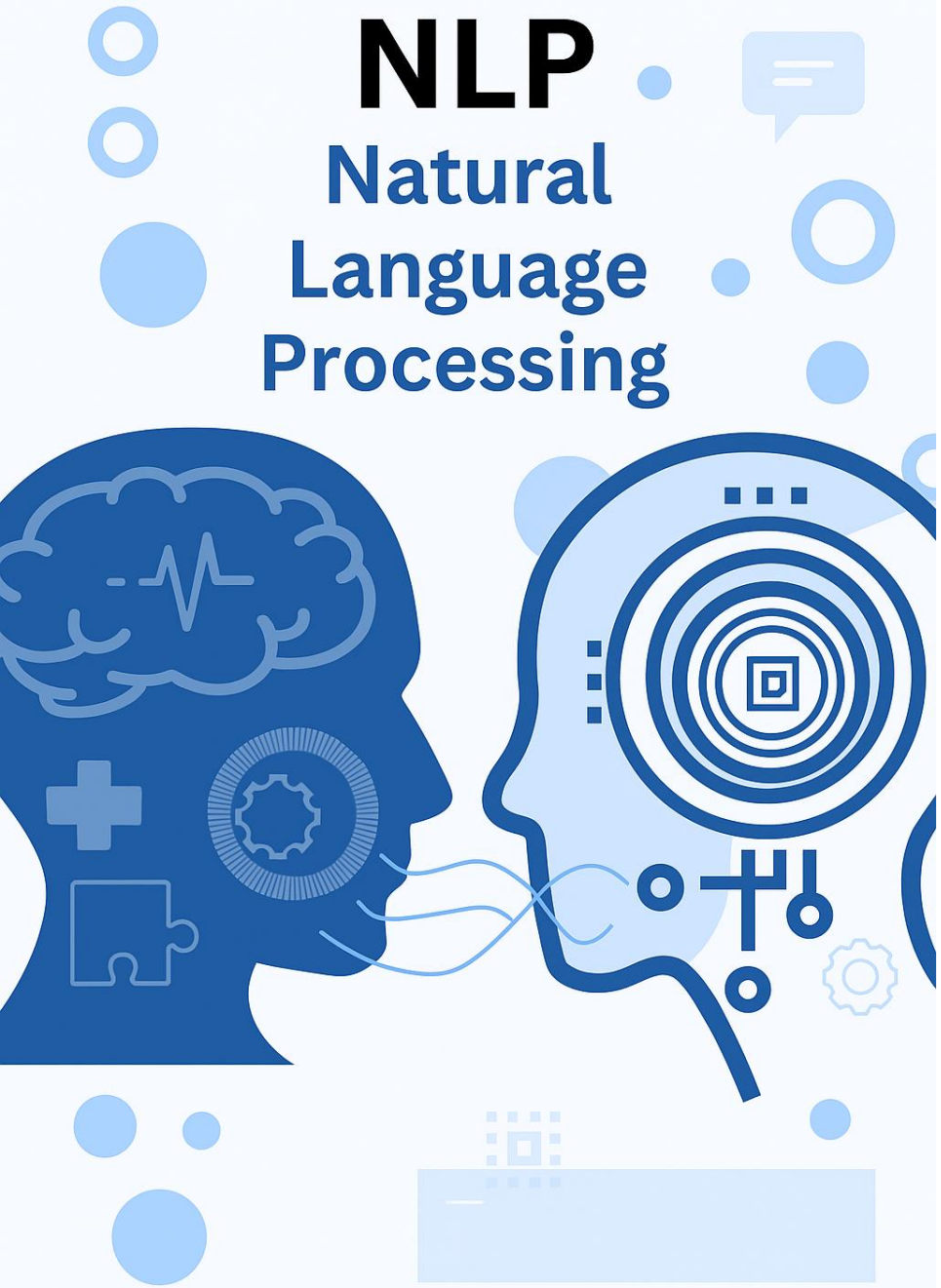


NATURAL LANGUAGE PROCESSING (NLP)

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

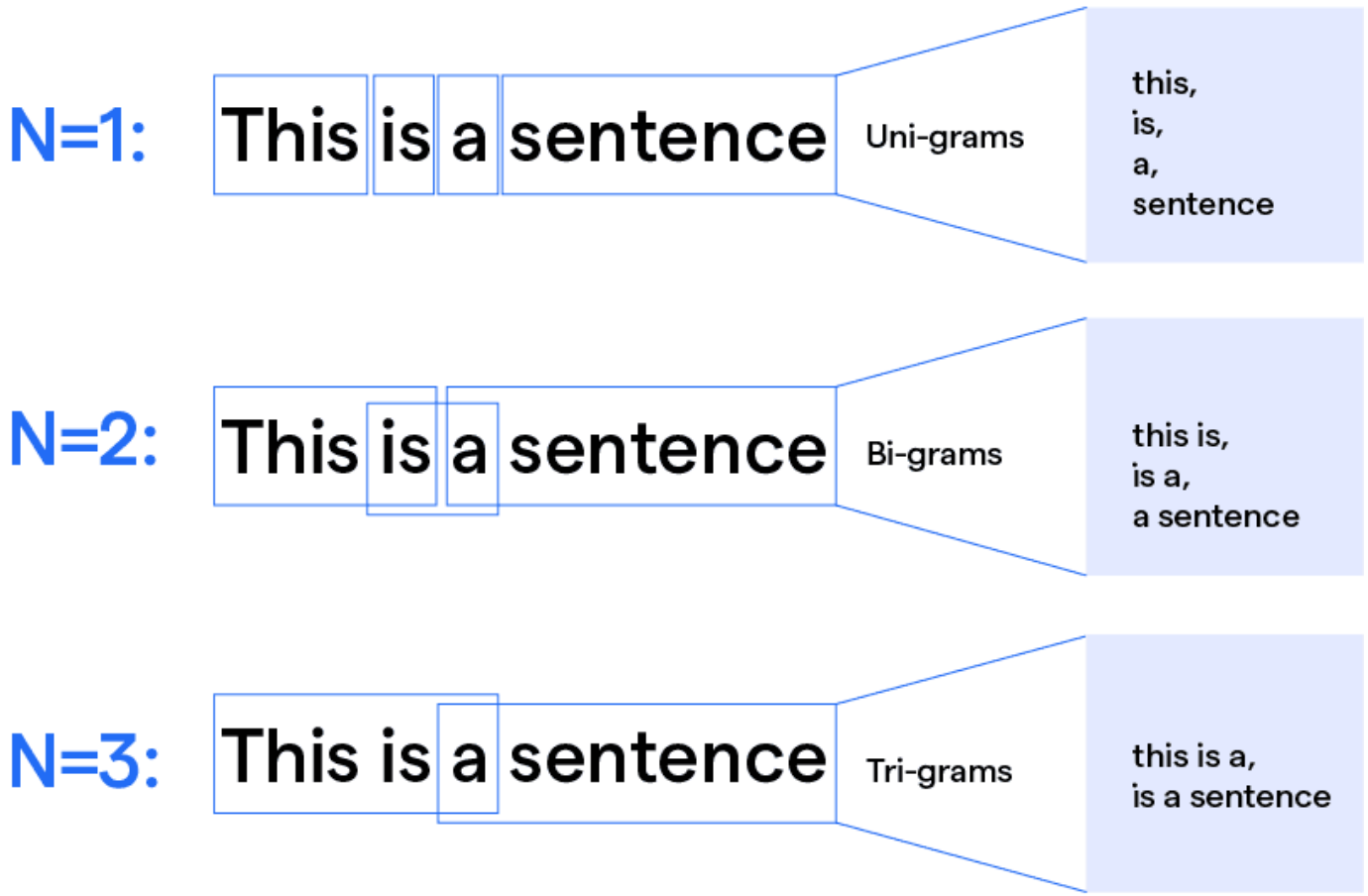
The following sequence has a non-zero probability of appearing in a text:

…..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

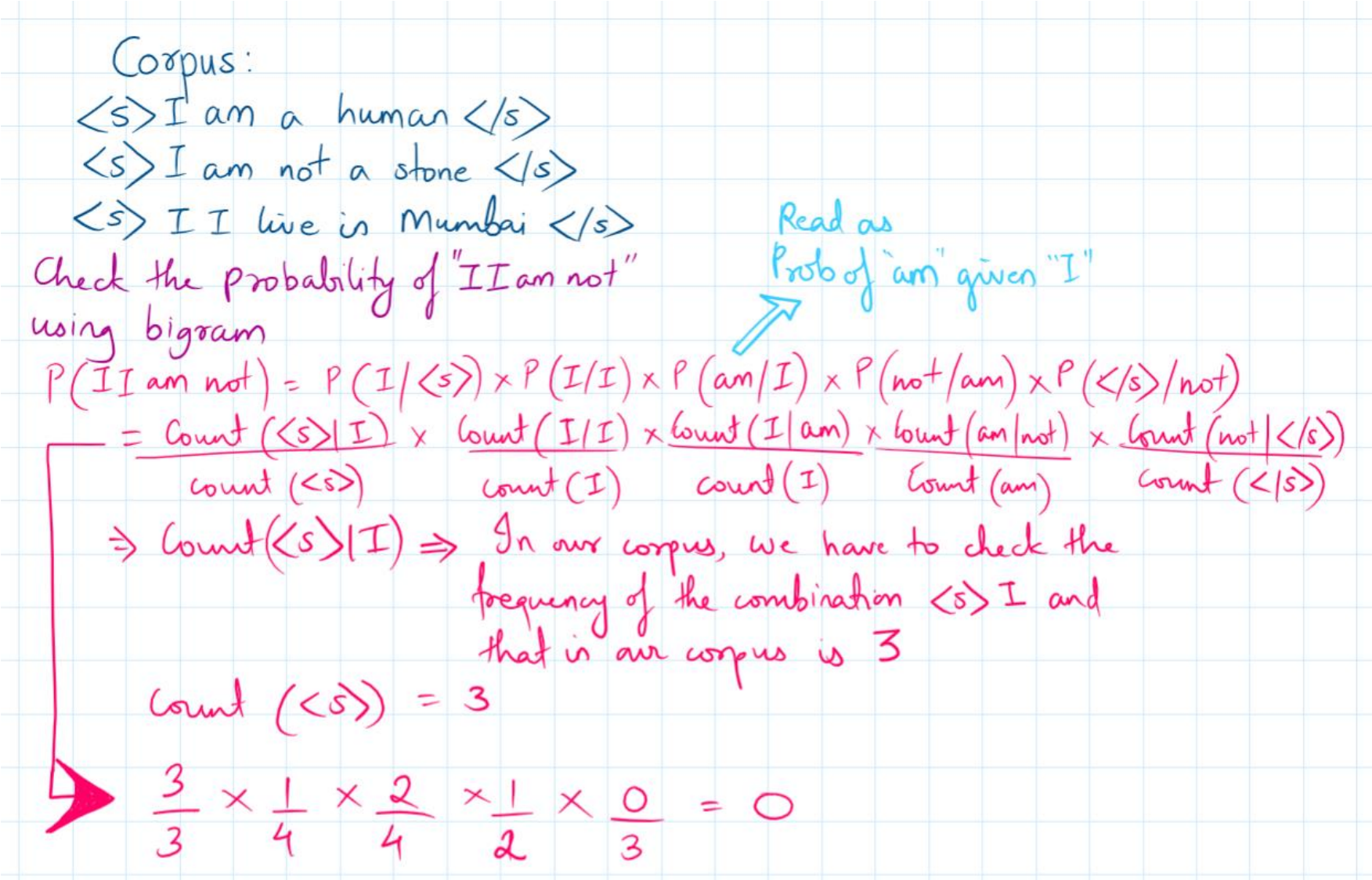

N=1:
This is a sentence
Uni-grams
this,
is,
a,
sentence
N=2:
This is a sentence
Bi-grams
this is,
is a,
a sentence
N=3:
This is a sentence
Tri-grams
this is a,
is a sentence

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**

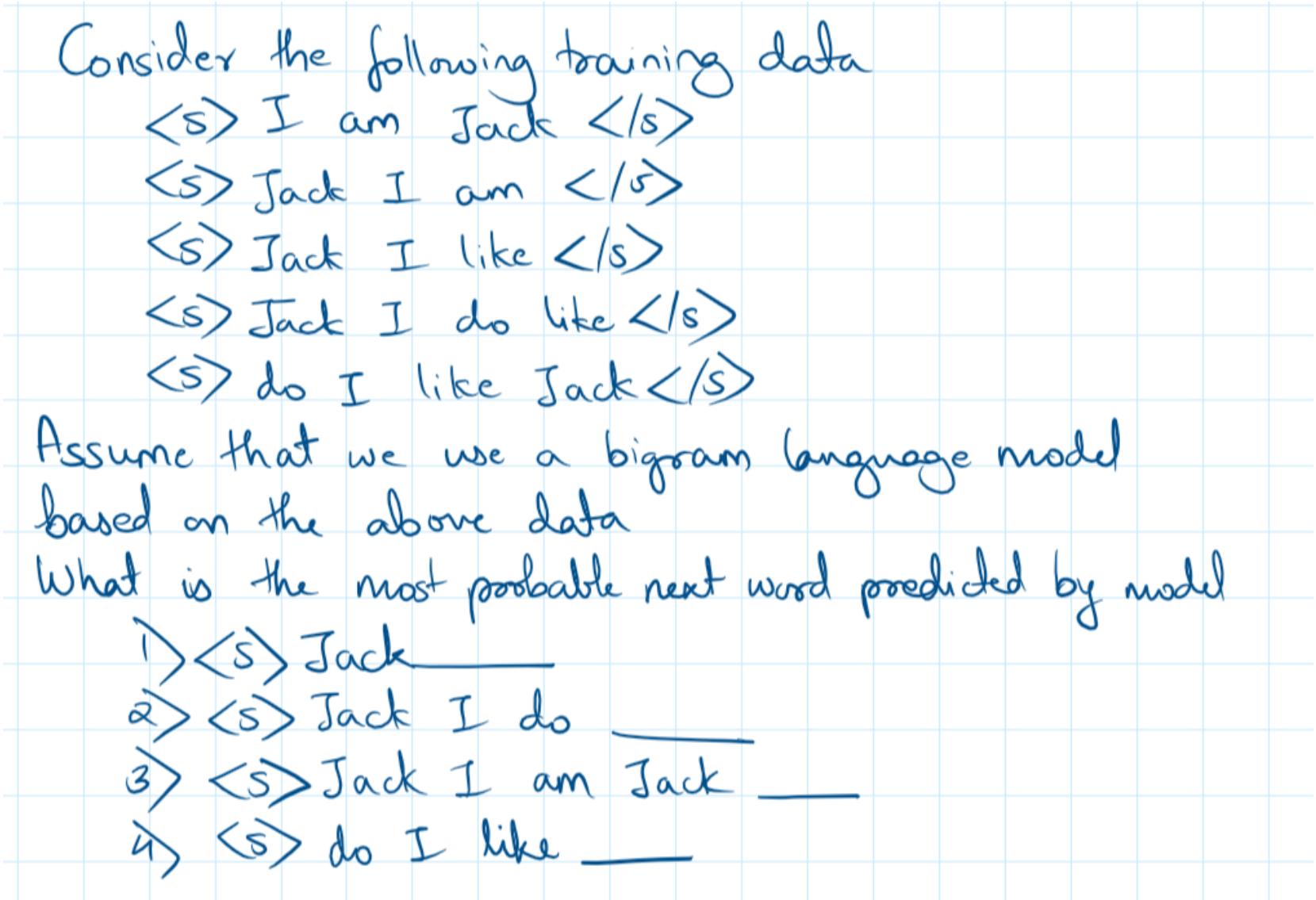
USING N-GRAMS TO PREDICT THE PROBABILITY OF A SENTENCE

For every sentence, we put <s> and </s> at the beginning and the end respectively. This denote the start and the end of the sentence.



USING N-GRAMS TO PREDICT THE NEXT WORD

For every sentence, we put <s> and </s> at the beginning and the end respectively. This denote the start and the end of the sentence.



USING N-GRAMS TO PREDICT THE NEXT WORD

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

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

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

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

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

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

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

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

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

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

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

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

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

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

USING N-GRAMS TO PREDICT THE NEXT WORD

1) Jack ____

⇒ P(something | Jack) = In our calculated probabilites we got 2 probabilities

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

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

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

2) Jack I do ____

P(something | do) → $P(I \mid do) = 1/2$, $P(like \mid do) = 1/2$ } The answer is both (I) and (like)