1. What is Bag of Words (BoW) in NLP?

Bag of Words is a method of converting text (sentences or documents) into numbers (vectors), so that machine learning models can understand it.

Key idea:

- We count how many times each word appears in the text.
- We ignore grammar and word order we only care about word frequency.

Example:

Let's say we have two simple sentences:

- Sentence 1: "I like NLP"
- Sentence 2: "I like AI"

Step 1: Create a vocabulary (list of all unique words):

```
CSS (Copy '2 Edit ["I", "like", "NLP", "AI"]
```

Step 2: Represent each sentence as a vector:

Word	Sentence 1	Sentence 2	
	1	1	
like			
NLP		0	
Al	0	1	

So,

- Sentence 1 → [1, 1, 1, 0]
- Sentence 2 → [1, 1, 0, 1]

2. What is Cosine Similarity?

Cosine Similarity measures how similar two vectors are, by measuring the angle between them.

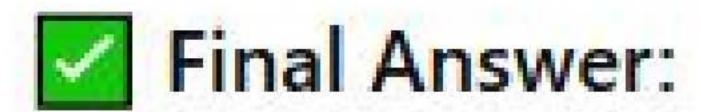
- If the angle is small → vectors are similar.
- Cosine similarity value ranges from -1 to 1.
 - 1 → exactly the same direction
 - 0 → completely different
 - -1 → opposite direction (rare in BoW cases)

Formula:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\| \times \|B\|}$$

Where:

- ullet $A\cdot B$ is the dot product of the two vectors.
- ullet $\|A\|$ is the magnitude (length) of vector A.



Cosine similarity ≈ 0.666

This means the sentences are fairly similar, but not identical.

Summary:

Concept	Description	
Bag of Words	Converts text into word-count vectors	
Cosine Similarity	Measures similarity between vectors	

- Sentence A: "The cat sat on the mat."
- Sentence B: "The dog sat beside the mat."



Step 1: Preprocess the text

We'll lowercase and remove punctuation (basic cleaning):

- . A: "the cat sat on the mat"
- B: "the dog sat beside the mat"

Step 2: Create Vocabulary

Make a list of all unique words from both sentences:

```
(5) Copy to Edit
CSS
["the", "cat", "sat", "on", "mat", "dog", "beside"]
```

(We skip duplicate words like "the" and "mat" and sort alphabetically just for clarity.)

Step 3: Create Bag-of-Words Vectors

Word	A ("cat sentence")	B ("dog sentence")	
the	2	2	
cat	1	0	
sat	1	1	
on	1	0	
mat	1	1	
dog	0	1	
beside	0	1	

So, the vectors are:

A = [2, 1, 1, 1, 1, 0, 0]

B = [2, 0, 1, 0, 1, 1, 1]

Step 4: Cosine Similarity Calculation

Dot Product:

$$(2 \times 2) + (1 \times 0) + (1 \times 1) + (1 \times 0) + (1 \times 1) + (0 \times 1) + (0 \times 1) + (0 \times 1) = 4 + 0 + 1 + 0 + 1 + 0 + 0 = 6$$

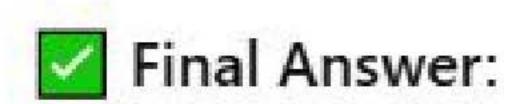
Magnitudes:

•
$$||A|| = \sqrt{(2^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2)} = \sqrt{(4+1+1+1+1)} = \sqrt{8} \approx 2.828$$

•
$$||B|| = \sqrt{(2^2 + 0^2 + 1^2 + 0^2 + 1^2 + 1^2 + 1^2)} = \sqrt{(4+1+1+1+1)} = \sqrt{8} \approx 2.828$$

Cosine Similarity:

$$\frac{6}{2.828 \times 2.828} = \frac{6}{8} = 0.75$$



Cosine Similarity = 0.75

This means the two sentences are quite similar — they share many common words, but also have some differences ("cat" vs "dog", etc).



Practice 1: Build BoW Vectors

Given these two sentences:

- S1: "I enjoy machine learning and deep learning."
- S2: "Deep learning is part of machine learning."

Step 1: Clean the Text

Lowercase and remove punctuation:

- S1: i enjoy machine learning and deep learning
- S2: deep learning is part of machine learning

Step 2: Build Vocabulary (unique words):

```
රා Copy % Edit
```

Step 3: Create Bag of Words Vectors

Word	S1	S2	
i	1	0	
enjoy	1	0	
machine	1	1	
learning	2	2	
and	1	0	
deep	1	1	
is	0	1	
part	0	1	
of	0	1	

So the vectors are:

• **S1**: [1, 1, 1, 2, 1, 1, 0, 0, 0]

• S2: [0, 0, 1, 2, 0, 1, 1, 1, 1]

Step 4: Cosine Similarity

Dot Product:

$$(1\times 0) + (1\times 0) + (1\times 1) + (2\times 2) + (1\times 0) + (1\times 1) + (0\times 1) + (0\times 1) + (0\times 1) + (0\times 1) = 0 + 0 + 1 + 4 + 0 + 1 + 0 + 0 + 0 = 6$$

Magnitudes:

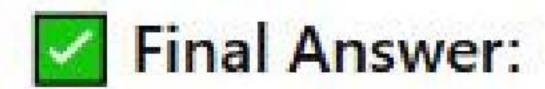
$$||S1|| = \sqrt{(1^2 + 1^2 + 1^2 + 2^2 + 1^2 + 1^2)} = \sqrt{(1 + 1 + 1 + 4 + 1 + 1)} = \sqrt{9} = 3$$

$$||S2|| = \sqrt{(1^2 + 2^2 + 1^2 + 1^2 + 1^2 + 1^2)} = \sqrt{(1 + 4 + 1 + 1 + 1 + 1)} = \sqrt{9} = 3$$

(Note: Only non-zero values contribute)

Cosine Similarity:

$$\frac{6}{3 \times 3} = \frac{6}{9} = 0.666$$



Cosine Similarity = 0.666

This means the two sentences are moderately similar, mainly because of shared terms like "machine",

"learning", and "deep".