TOPIC SEGMENTATION USING SENTENCE EMBEDDINGS

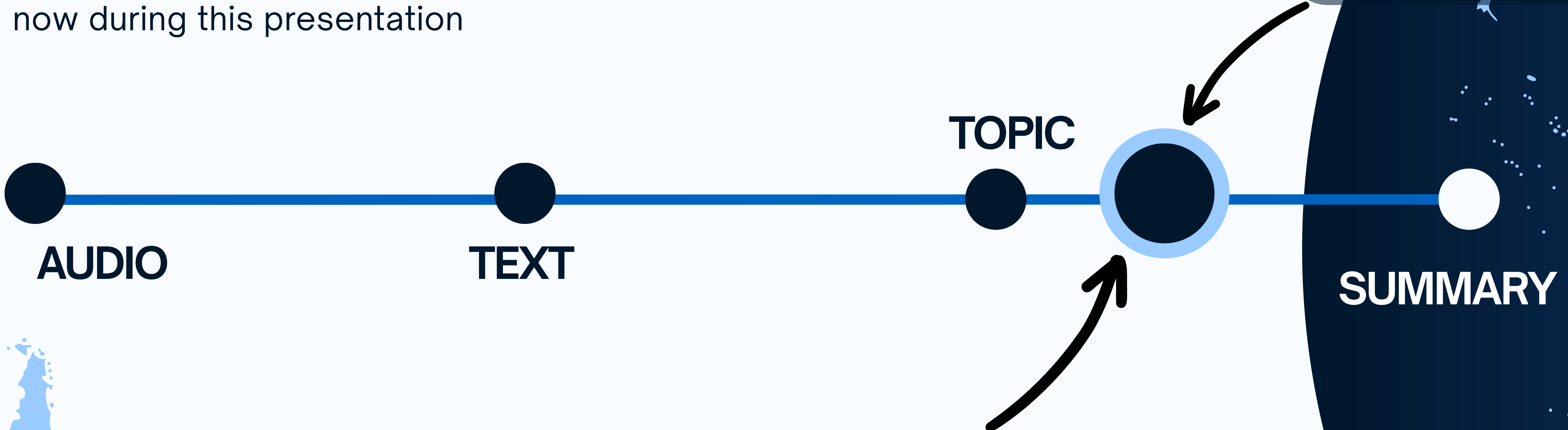
Detecting topic shifts using Semantic similarity

Goutham Gourabathuni

G GOUTHAM

ROADMAP

Our location in the whole project pipeline right now during this presentation



Topic segmentation using semantic similarity.

SOURCE CODE

76% progress covered

G GOUTHAM

OBJECTIVES OF THIS PRESENTATION

Important OBJECTIVES

1) Sentence Embeddings

2) Cosine Similarity

3) Threshold - based Topic change

MAIN OBJECTIVE

4) Detecting Topic shifts using semantic similarity

1

SENTENCE EMBEDDINGS

Numerical vectors representing Sentences

G GOUTHAM

SENTENCE EMBEDDINGS

Definition, Need and creation of embeddings

What Are Sentence Embeddings?

Sentence embeddings are fixed-length numerical vectors that represent the meaning of an entire sentence.

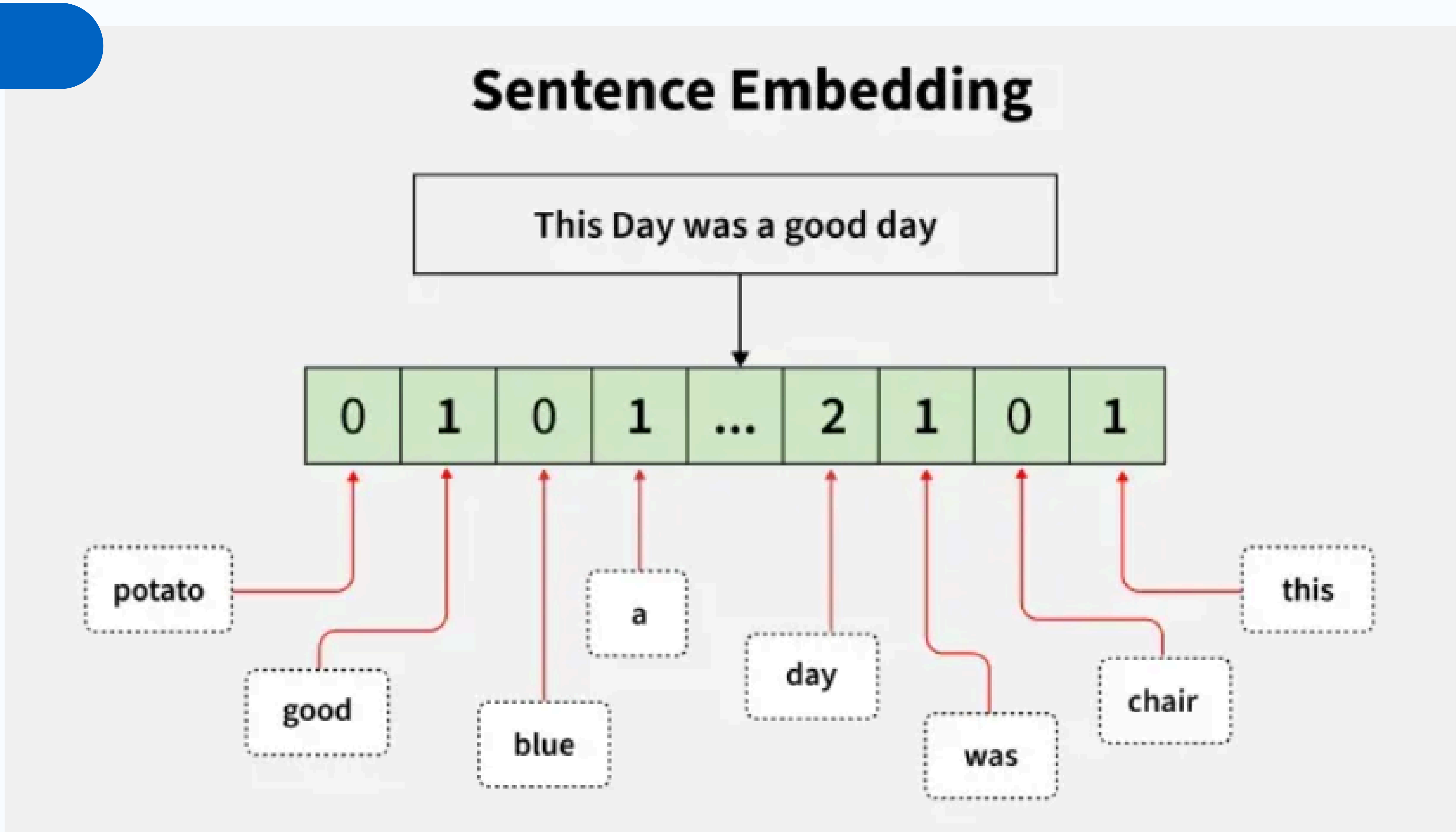
Why We Need Sentence Embeddings?

Computers can’t understand the text. Therefore, Embeddings act as a bridge for easy machine understanding

How Sentence Embeddings Are Created?

- Sentence is tokenized
- Passed through a transformer encoder
- Token representations are pooled
- Final vector = sentence embedding

G GOUTHAM

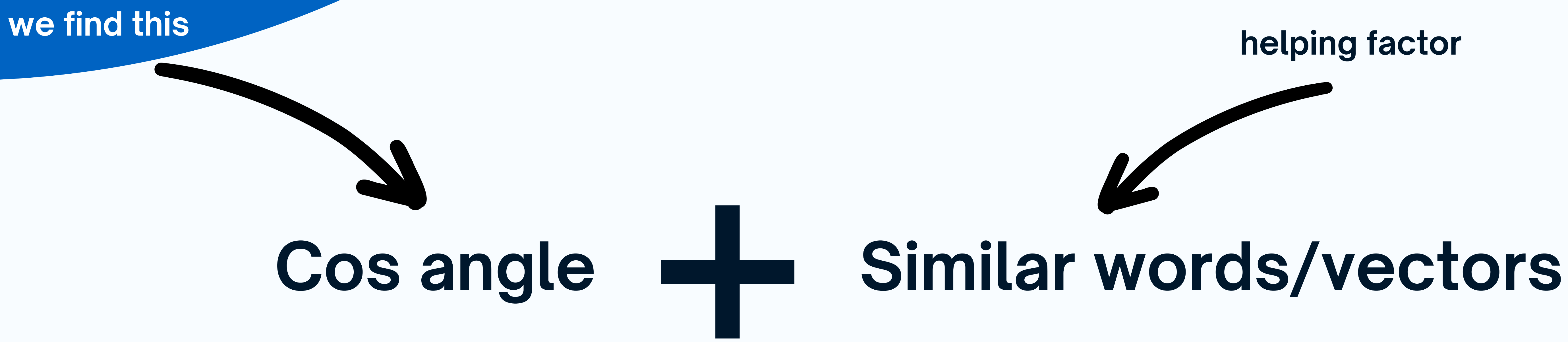


In this image, each word of the sentence is transformed into numeric vectors and zero is imparted to words which are not present in the sentence.

Sentence transformer used to generate these embeddings:- SBERT

2

COSINE SIMILARITY



we find this

helping factor

Cos angle + Similar words/vectors

G GOUTHAM

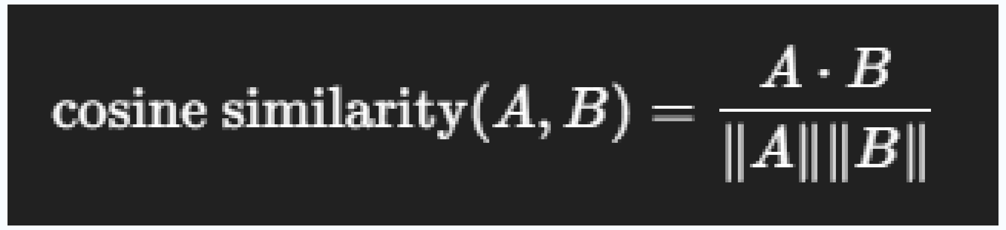
WHAT IS COSINE SIMILARITY ??

Cosine Similarity (Definition):- Cosine similarity is a measure of how similar two vectors are based on the angle between them, not their length.

If two sentences mean similar things → their vectors point in almost the same direction.

If they talk about different topics → their vectors point in different directions.

Mathematical Definition



$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

Where:

- **$A \cdot B$ → dot product (overlap of meaning)**
- **$||A||$ and $||B||$ → vector lengths**

Value	Meaning
1	Identical meaning
~0.7–0.9	Strongly related
~0.4–0.6	Weakly related
~0	Unrelated

3

THRESHOLD BASED TOPIC CHANGE

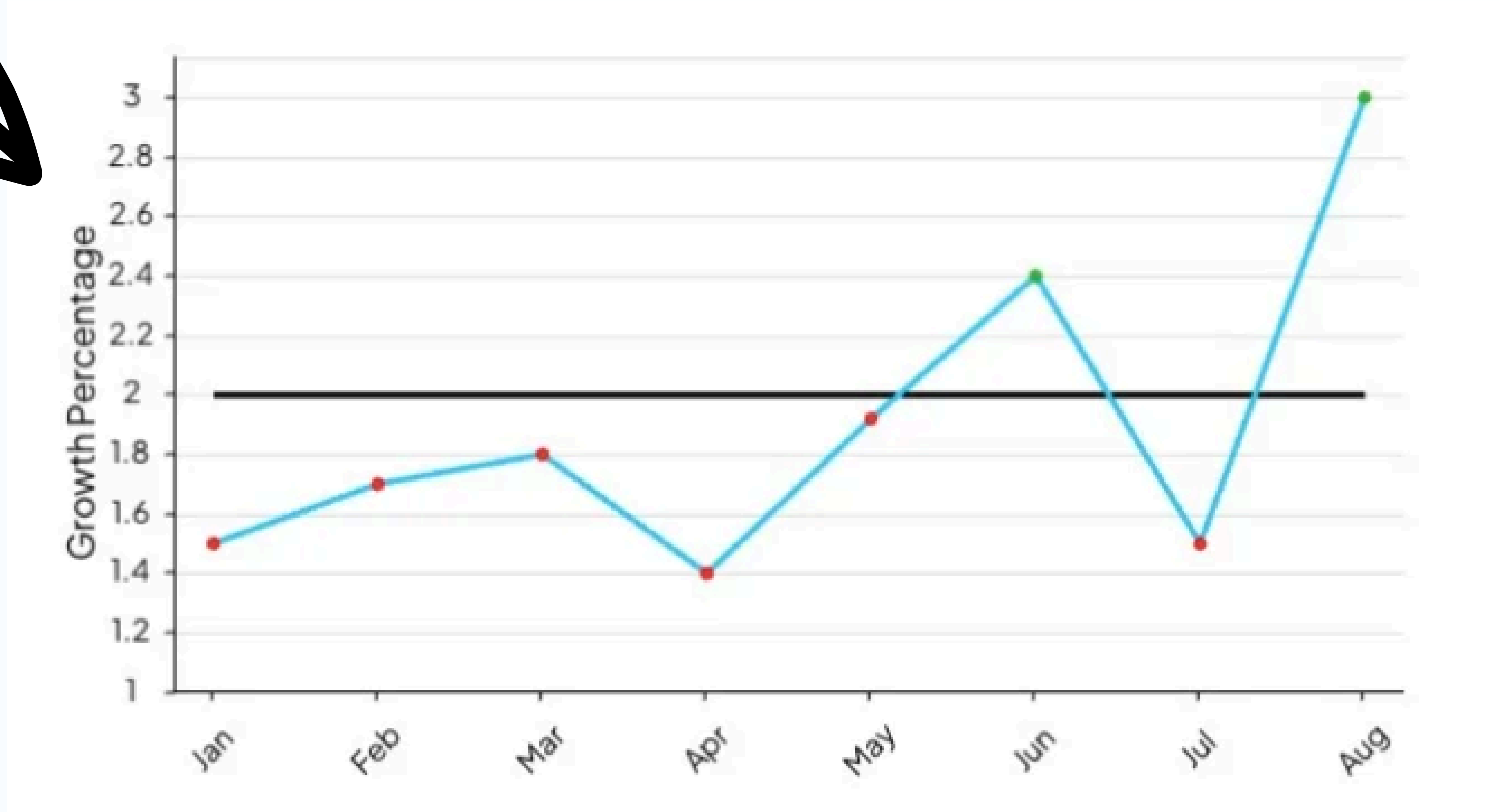
cutoff

Threshold + Change of topics

WHAT IS THRESHOLD ??

Threshold (Definition):- A threshold is a predefined cutoff value used to make a decision.

2 is the threshold here



Example

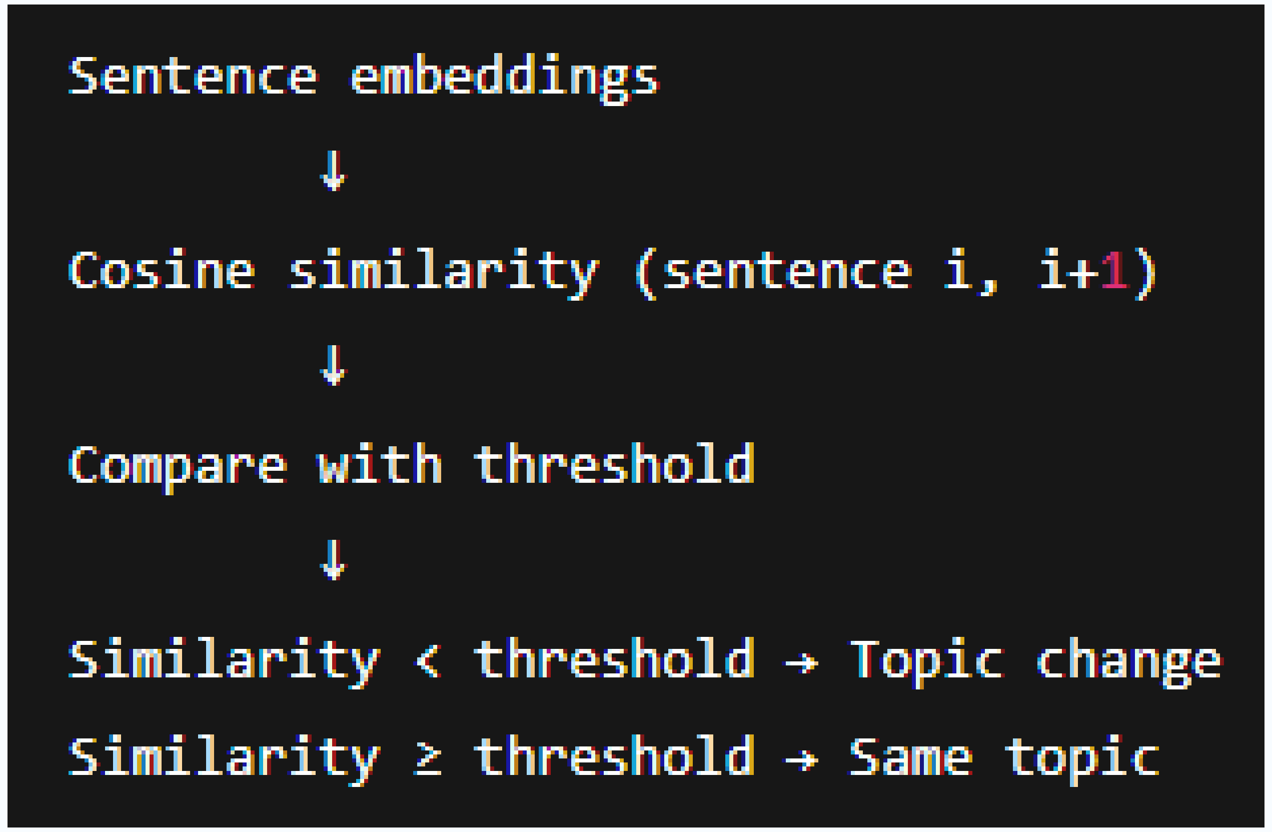
- Pass mark = 40
  - Score ⩾ 40 → Pass
  - Score < 40 → Fail

The number 40 is the threshold.

G GOUTHAM

WHAT IS THRESHOLD-BASED TOPIC CHANGE?

When a point comes below the threshold of the graph, it means the topic is different and has experienced a change compared to the previous one.



```
Sentence embeddings
        ↓
Cosine similarity (sentence i, i+1)
        ↓
Compare with threshold
        ↓
Similarity < threshold → Topic change
Similarity ≥ threshold → Same topic
```

When a point comes above the threshold of the graph, it means the topic has not changed.

> “In topic segmentation using semantic similarity, a threshold is applied to cosine similarity scores between consecutive sentences.”

4

DETECTING TOPIC SHIFTS USING SEMANTIC SIMILARITY

we find this

helping factor

Topic shifts + Semantic similarity

G GOUTHAM

WHAT ARE TOPIC SHIFTS ??

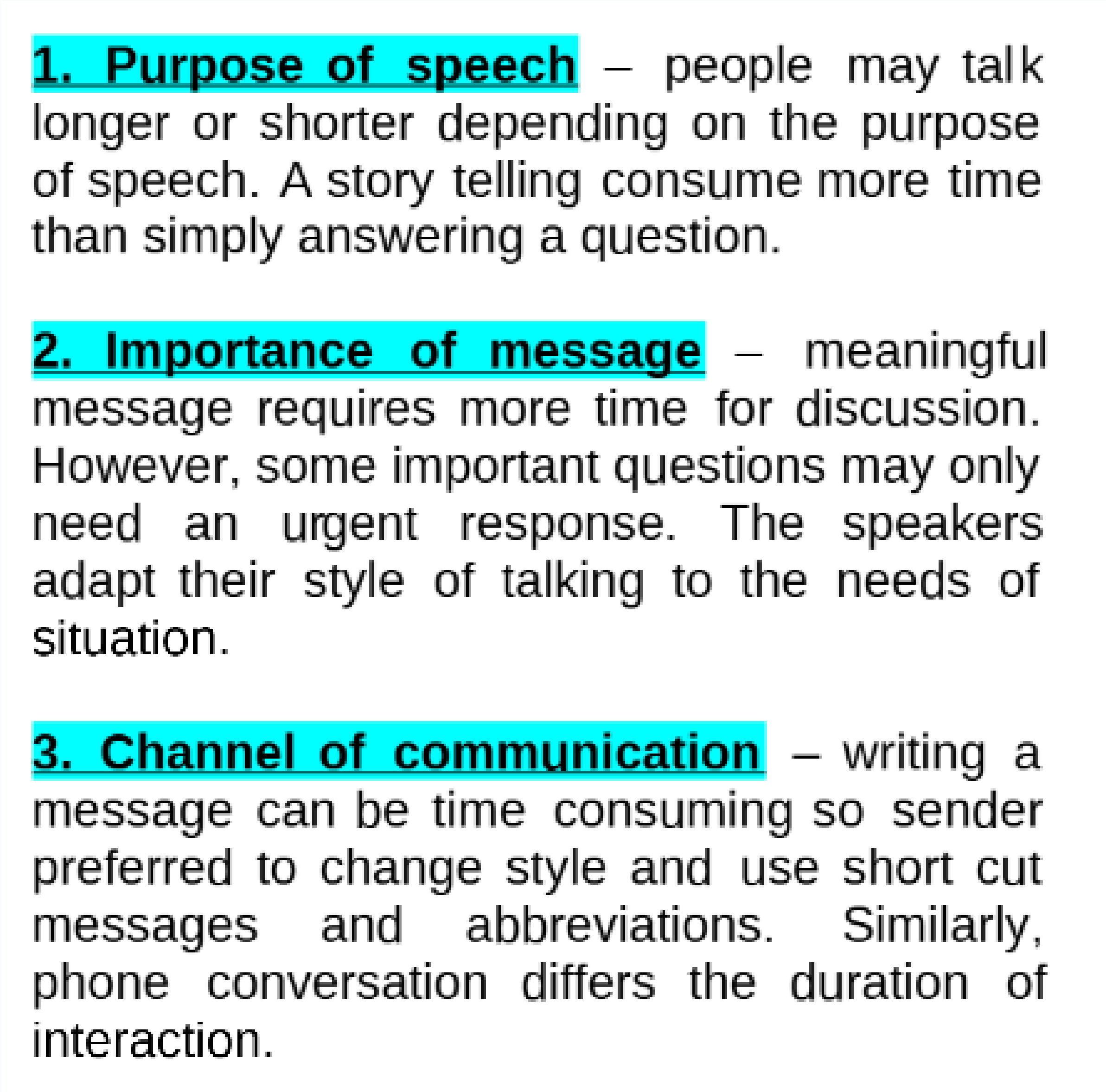
Topic Shifts (Definition):- A topic shift is a point in a transcript where the conversation changes focus

This shift marks the boundary between two different (meaningful) semantic topics. “Separation point between 2 paras”

Example:-

“Today, AI is transforming healthcare…

Next let’s talk about venture capital investments.”



1. Purpose of speech – people may talk longer or shorter depending on the purpose of speech. A story telling consume more time than simply answering a question.

2. Importance of message – meaningful message requires more time for discussion. However, some important questions may only need an urgent response. The speakers adapt their style of talking to the needs of situation.

3. Channel of communication – writing a message can be time consuming so sender preferred to change style and use short cut messages and abbreviations. Similarly, phone conversation differs the duration of interaction.

GOUTHAM

WHAT IS SEMANTIC SIMILARITY ??

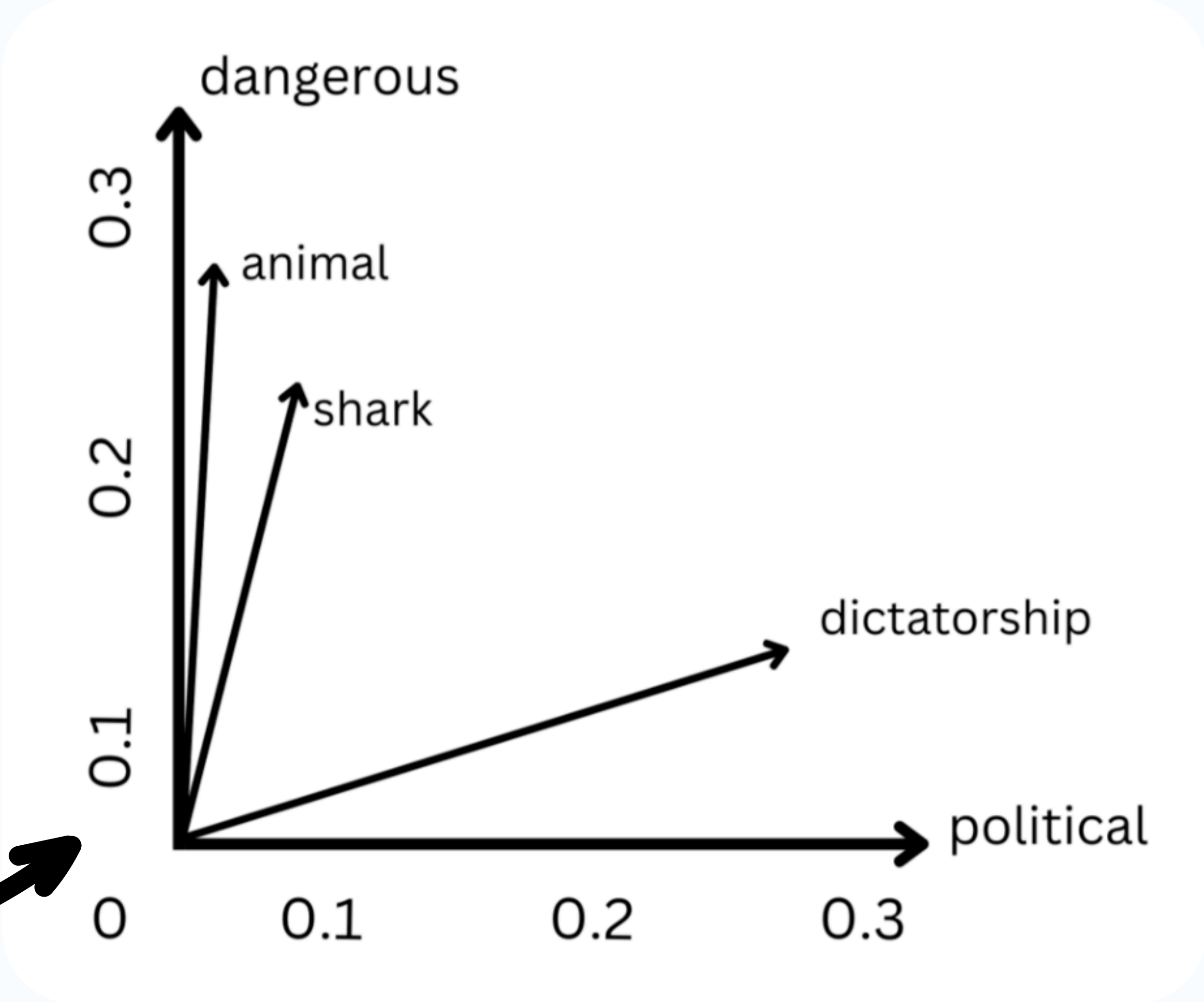
Semantic Similarity (Definition):- Semantic similarity measures how similar two pieces of text are in meaning, not just in words.

Instead of exact word matches, semantic similarity relies on vector representations from language models (e.g., SBERT, BERT)

In Vector Space

“similar meaning → vectors are close in space

dissimilar meaning → vectors are far apart”



The vector space

G GOUTHAM

EXECUTION

Let’s Detect Topic Shifts Using Semantic Similarity

1 Load a model and
Convert each sentence → sentence embedding

2 Compute cosine similarity between embeddings of consecutive sentences

3 If similarity falls below a threshold → it likely indicates a topic shift

SOURCE CODE ●

click for testing the execution.

THANK YOU

☎ +91 6309971208

✉ gouthamgourabathuni@gmail.com

Goutham Gourabathuni