Cosine Similarity

Applications in Artificial Intelligence

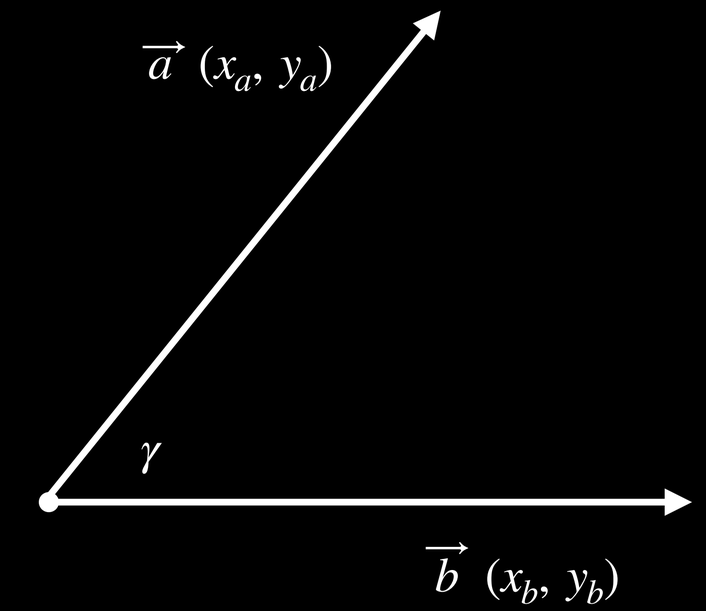
David Oniani

The idea is simple!

It is all about the dot product!

There are 4 steps.

All of them are simple!



(1) $\overrightarrow{a} \cdot \overrightarrow{b} = x_a \times x_b + y_a \times y_b$

(2) $\overrightarrow{a} \cdot \overrightarrow{b} = |\overrightarrow{a}| \times |\overrightarrow{b}| \times \cos\gamma$

(3) $x_a \times x_b + y_a \times y_b = |\overrightarrow{a}| \times |\overrightarrow{b}| \times \cos\gamma$

(4) $\cos\gamma = \dfrac{x_a \times x_b + y_a \times y_b}{|\overrightarrow{a}| \times |\overrightarrow{b}|}$

$$0 \leq |\cos\gamma| \leq 1$$

As long as they are vectors…

Works for Text Documents

How cool is that?

This is a sentence

$$\begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Counts of letters “a” and “e”

$$\approx 0.997$$

This is another sentence

$$\begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

Abstract > Vectorize > Apply

Vectorization Techniques

There are a lot of them!

- Count Vectorization
- tf-idf
- Word2Vec
- USE (Universal Sentence Encoder)
- BERT
- BioBERT
- Custom Vectorization Techniques

Thank you!