Cosine Similarity

What is Cosine Similarity?

Cosine similarity is a metric used to measure the similarity of two vectors. Specifically, it measures the similarity in the direction or orientation of the vectors ignoring differences in their magnitude or scale. Both vectors need to be part of the same inner product space, meaning they must produce a scalar through inner product multiplication. The similarity of two vectors is measured by the cosine of the angle between them.

How to calculate Cosine Similarity

We define cosine similarity mathematically as the dot product of the vectors divided by their magnitude. For example, if we have two vectors, A and B, the similarity between them is calculated as:

$$similarity(A, B) = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

where

- θ is the angle between the vectors,
- $A \cdot B$ is dot product between A and B and calculated as $A \cdot B = A^T B = \sum_{i=1}^{n} A_i B_i = A_1 B_1 + A_2 B_2 + \ldots + A_n B_n$,
- $\|A\|$ represents the L2 norm or magnitude of the vector which is calculated as $\|A\| = \sqrt{A_1^2 + A_1^2 \ldots A_1^n}$.

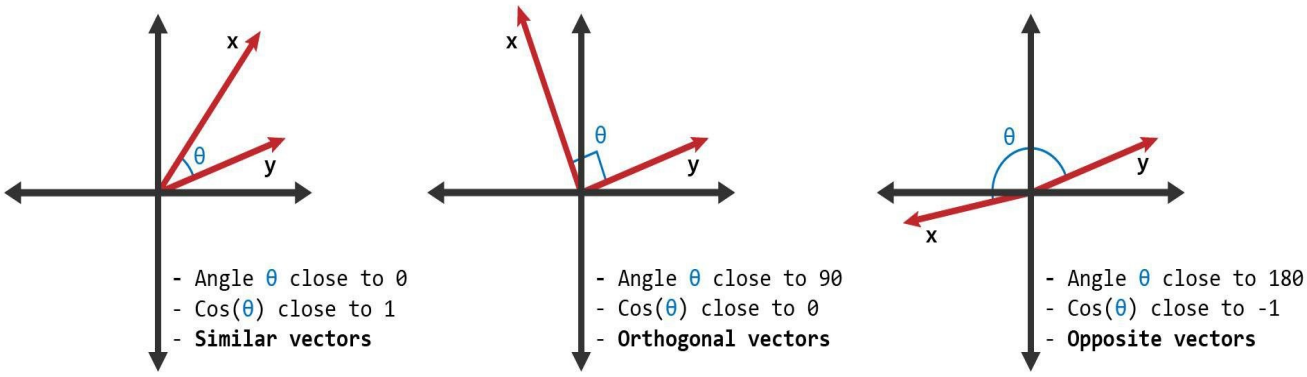
The similarity can take values between -1 and +1. Smaller angles between vectors produce larger cosine values, indicating greater cosine similarity. For example:

•When two vectors have the same orientation, the angle between them is 0, and the cosine similarity is 1.

•Perpendicular vectors have a 90-degree angle between them and a cosine similarity of 0.

•Opposite vectors have an angle of 180 degrees between them and a cosine similarity of -1.

Here's a graphic showing two vectors with similarities close to 1, close to 0, and close to -1.



Applications

Cosine similarity is beneficial for applications that utilize sparse data, such as word documents, transactions in market data, and recommendation

systems because cosine similarity ignores 0-0 matches. Counting 0-0 matches in sparse data would inflate similarity scores. Another commonly used metric that ignores 0-0 matches is Jaccard Similarity.
Cosine Similarity is widely used in Data Science and Machine Learning applications. Examples include measuring the similarity of:

1. Documents in natural language processing
2. Movies, books, videos, or users in recommendation systems
3. Images in computer vision

Numerical Example

Suppose that our goal is to calculate the cosine similarity of the two documents given below.

1. **Document 1 = 'the best data science course'**
2. **Document 2 = 'data science is popular'**

	the	best	data	science	course	is	popular
D1	1	1	1	1	1	0	0
D2	0	0	1	1	0	1	1

- $D1 = [1, 1, 1, 1, 1, 0, 0]$
- $D2 = [0, 0, 1, 1, 0, 1, 1]$

After creating a word table from the documents, the documents can be represented by the following vectors:

Using these two vectors we can calculate cosine similarity. First, we calculate the dot product of the vectors:

$$D1 \cdot D2 = 1 \times 0 + 1 \times 0 + 1 \times 1 + 1 \times 1 + 1 \times 0 + 0 \times 1 + 0 \times 1 = 2$$

Second, we calculate the magnitude of the vectors:

$$\|D1\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2} = \sqrt{5}$$

$$\|D2\| = \sqrt{0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 1^2} = \sqrt{4}$$

Finally, cosine similarity can be calculated by dividing the dot product by the magnitude

$$similarity(D1, D2) = \frac{D1 \cdot D2}{\|D1\|\|D2\|} = \frac{2}{\sqrt{5}\sqrt{4}} = \frac{2}{\sqrt{20}} = 0.44721$$

The angle between the vectors is calculated as:

Cosine Similarity is **0.44721.**

Angle is **63.43**

$$cos(\theta) = 0.44721$$

$$\theta = \arccos(0.44721) = 63.435$$

The cosine similarity is not close to 1, its 0.44721 shows the two Sentences are not similar, BUT they are around 44% similar.

The angle between them is larger.(63.435)

Python Example
https://colab.research.google.com/drive/1axfHgSWPNLLunnjUwCpNTVKfLiS24Sa9?usp=sharing

Read more
https://www.geeksforgeeks.org/cosine-similarity/
https://towardsdatascience.com/understanding-cosine-similarity-and-its-application-fd42f585296a
https://en.wikipedia.org/wiki/Cosine_similarity