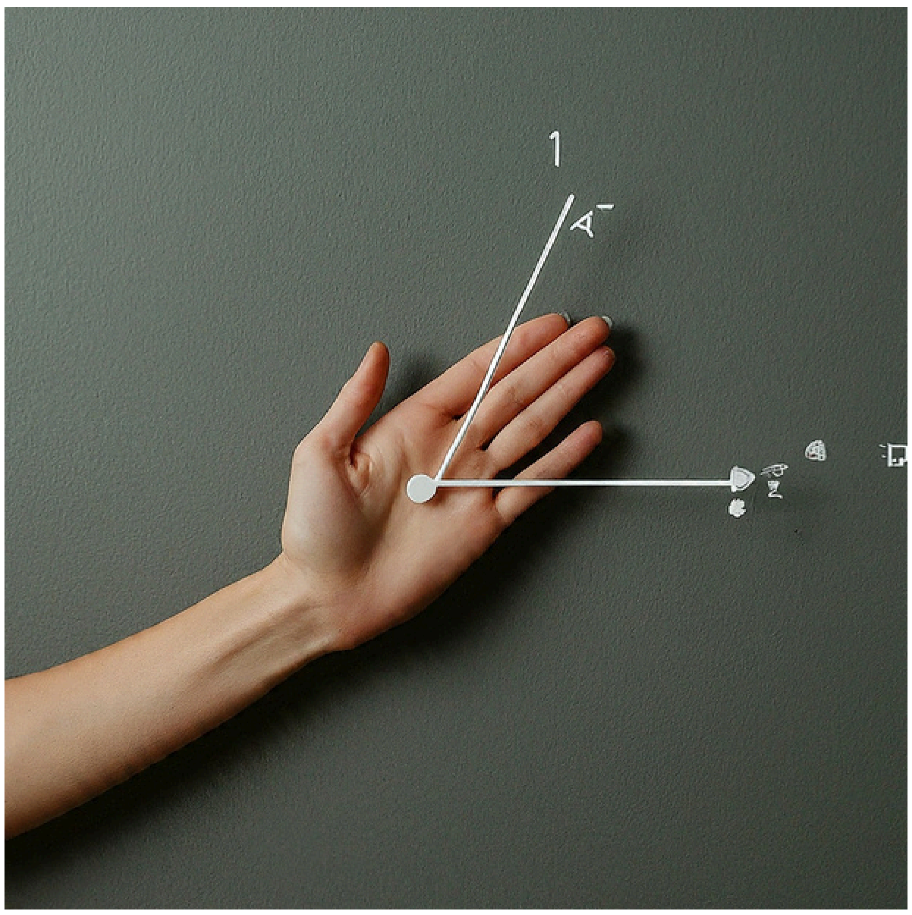
A Beginner's Guide to Cosine Similarity

What is Cosine Similarity?

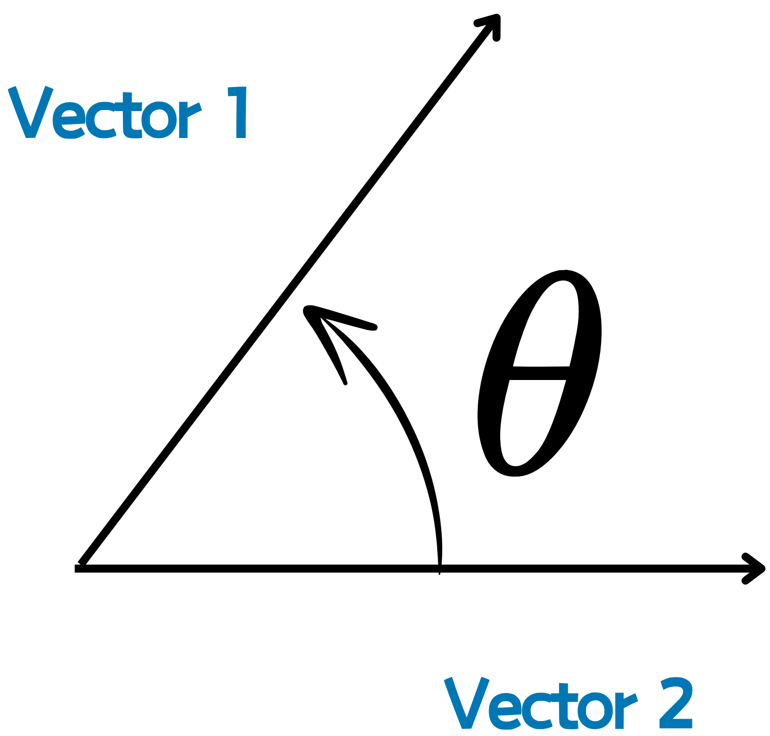
- Cosine similarity is a metric used to measure the similarity between two text documents or snippets.
- It is widely used in search engines, plagiarism detection, and recommendation systems to compare the content of texts and determine how closely they are related.
- This article explains the concept of cosine similarity, its calculation, and its applications in various domains, using simple examples for clarity.



Defining Cosine Similarity

Definition:

- Cosine similarity is a measure of similarity between two non-zero vectors in an inner product space. It is defined as the cosine of the angle between the vectors.
- In the context of textual data, these vectors represent the frequency or importance of terms within the documents.



Cosine Similarity = $Cos\,(\theta)$

How Cosine Similarity Works

How Cosine Similarity Works

Vector Representation:

- Each document is represented as a vector in a multi-dimensional space. The dimensions correspond to the unique terms in the corpus (all documents combined).
- Common methods to convert text to vectors include term frequency (TF) and term frequency-inverse document frequency (TF-IDF).

How Cosine Similarity Works

Formula:

The cosine similarity **sim(A,B)** between two vectors **A** and **B** is given by:

$$\mathrm{sim}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

Where:

- A . B is the dot product of vectors A and B
- ||A|| and ||B|| are the magnitudes (or norms) of vectors A and B

Interpretation:

- **The value ranges from 0 to 1.**
- **1 indicates that the vectors are identical.**
- **0 indicates that the vectors are orthogonal (no similarity).**
- **Values between 0 and 1 indicate varying degrees of similarity.**

Applications of Cosine Similarity

1. **Search Engines:**
   - When a user enters a query, the search engine represents the query and documents as vectors.
   - Cosine similarity is used to rank documents based on their relevance to the query.
   - Higher cosine similarity scores indicate more relevant documents.
2. **Plagiarism Detection:**
   - Documents are compared to identify similar content.
   - High cosine similarity between two documents suggests potential plagiarism.
   - This method can detect paraphrased or partially copied content.
3. **Recommendation Systems:**
   - Used to recommend items based on textual descriptions or user reviews.
   - Items with high cosine similarity to a user's preferences are recommended.
   - Example: Suggesting articles or products similar to those previously interacted with by the user.

Example Calculations

Example 1: Simple Document Comparison

Consider two short documents:

- **Document 1: "Data science is fun"**
- **Document 2: "I love data science"**

Tokenization and Vectorization:

- Unique terms: [data, science, is, fun, I, love]
- Document vectors (binary representation):
- Document 1: [1, 1, 1, 1, 0, 0]
- Document 2: [1, 1, 0, 0, 1, 1]

Dot Product and Magnitude:

- **Dot product:**

$$1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 1 = 2$$

- **Magnitude of Document 1:**

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

Example Calculations

Example 1: Simple Document Comparison

- **Magnitude of Document 2:**

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

- **Cosine Similarity:**

$$\text{sim}(\text{Doc1}, \text{Doc2}) = \frac{2}{2 \cdot 2} = \frac{2}{4} = 0.5$$

Example Calculations

Example 2: Simple Document Comparison

Consider two documents:

- Document 1: "Artificial intelligence and machine learning are evolving fields"
- Document 2: "Machine learning is a subset of artificial intelligence"

Tokenization and Vectorization:

- Unique terms: [artificial, intelligence, and, machine, learning, are, evolving, fields, is, subset, of]
- Document vectors (binary representation):
- Document 1: [1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0]
- Document 2: [1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1]

Dot Product and Magnitude:

- **Dot product:**

  Dot product: $1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 0 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 1 = 4$

- **Magnitude of Document 1:**

$$\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2} = \sqrt{8} \approx 2.83$$

Example Calculations

Example 2: Simple Document Comparison

- **Magnitude of Document 2:**

$$\sqrt{1^2 + 1^2 + 0^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2} = \sqrt{7} \approx 2.65$$

- **Cosine Similarity:**

$$\mathrm{sim}(\mathrm{Doc1}, \mathrm{Doc2}) = \frac{4}{2.83 \cdot 2.65} \approx \frac{4}{7.5} \approx 0.53$$

Example Calculations

Example 3: Plagiarism Detection

Consider two documents:

- **Document 1: "The quick brown fox jumps over the lazy dog"**
- **Document 2: "The quick brown fox leaps over the lazy dog"**

Tokenization and Vectorization:

- Unique terms: [the, quick, brown, fox, jumps, over, lazy, dog, leaps]
- Document vectors (binary representation):
- Document 1: [1, 1, 1, 1, 1, 1, 1, 1, 0]
- Document 2: [1, 1, 1, 1, 0, 1, 1, 1, 1]

Dot Product and Magnitude:

- **Dot product:**

$$1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 0 \cdot 1 = 7$$

- **Magnitude of Document 1:**

$$\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2} = \sqrt{8} \approx \ 2.83$$

Example Calculations

Example 3: Plagiarism Detection

- **Magnitude of Document 2:**

$$\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{8} \approx \quad 2.83$$

- **Cosine Similarity:**

$$\text{sim}(\text{Doc1}, \text{Doc2}) = \frac{7}{2.83 \cdot 2.83} \approx \frac{7}{8} \approx 0.875$$

Summary

- Cosine similarity is a powerful and widely used metric for comparing textual data.
- Its applications in search engines, plagiarism detection, and recommendation systems highlight its importance in modern data science and information retrieval.
- By understanding and leveraging cosine similarity, we can enhance the performance of various systems that rely on

Special Thanks to ChatGPT
and Gemini for Content support