Cosine similarity

Cosine similarity is the cosine of the angle between the vectors; that is, it is the dot product of the vectors divided by the product of their lengths. It does not depend on the magnitudes of the vectors, but only on their angle.

The formula to find the cosine similarity between two vectors is –

$$\cos(x, y) = x \,.\, y / ||x|| * ||y||$$

where,

● x . y = product (dot) of the vectors ‘x’ and ‘y’.

● ||x|| and ||y|| = length of the two vectors ‘x’ and ‘y’.

● ||x|| * ||y|| = cross product of the two vectors ‘x’ and ‘y’.

In data analysis, cosine similarity is a measure of similarity between two non-zero vectors defined in an inner product space. Cosine similarity is the cosine of the angle between the vectors; that is, it is the dot product of the vectors divided by the product of their lengths. It follows that the cosine similarity does not depend on the magnitudes of the vectors, but only on their angle. The cosine similarity always belongs to the interval

[−1,1].

For example, two proportional vectors have a cosine similarity of 1, two orthogonal vectors have a similarity of 0, and two opposite vectors have a similarity of -1. In some contexts, the component values of the vectors cannot be negative, in which case the cosine similarity is bounded in

[0,1].

The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$

Given two n-dimensional vectors of attributes, A and B, the cosine similarity, cos(θ), is represented using a dot product and magnitude as

$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

Objective behind the algorithm:

Due to its simplicity in implementation and computational efficiency, the cosine similarity algorithm is extensively used in information retrieval. Moreover, it is capable of capturing the semantic similarity between documents accurately, even if they have distinct word choices or varying lengths.

Cosine similarity is a prevalent technique in information retrieval to prioritize documents according to their relevance to a user's query.

The main goal of using the cosine similarity algorithm is to obtain a dependable measure of similarity between documents. Such a measure can significantly enhance the precision and efficiency of information retrieval systems.

Advantages:

- The cosine similarity measure is frequently utilized in Information Retrieval due to its simplicity, efficiency, and effectiveness.

- This technique is highly robust, as it is insensitive to the order of terms in documents, making it appropriate for a broad range of document types and applications.

- Cosine similarity is adept at capturing the semantic similarity between documents, even in cases of varying lengths or word choices.

- By ranking and retrieving documents based on their relevance to a user's query, cosine similarity is a valuable tool in improving the precision and efficiency of IR systems.

- Additionally, cosine similarity scores are easily interpretable, and they can offer useful insights into the relationships between documents in a corpus.

- Combined with other techniques such as term weighting and query expansion, cosine similarity can enhance the performance of IR systems.

- The calculation of cosine similarity scores can be parallelized with ease, enabling the efficient processing of large-scale document collections.

- Moreover, cosine similarity is widely supported by open-source libraries and APIs, making it easily accessible to researchers and practitioners in the field.

Disadvantages:

- One limitation of cosine similarity is that it does not take into account the importance of rare terms in a document, which can affect the accuracy of the similarity scores and ranking of documents.

- Additionally, cosine similarity assumes that all dimensions in the vector space are equally significant, which is not always the case. Some terms may have a higher relevance to a particular document or query than others.

- The sensitivity of cosine similarity to document length can lead to higher similarity scores for longer documents, which may not always reflect their true similarity to the query.

- It may not be suitable for measuring similarity between documents in different domains or topics, as the vocabulary and term frequencies can vary significantly.

- The use of stop words and stemming can impact the accuracy of cosine similarity scores and document ranking, potentially resulting in incorrect results.

- In situations where there is a considerable overlap between the query terms and the terms in a document, cosine similarity may not accurately distinguish between relevant and irrelevant documents.
- Cosine similarity is a bag-of-words approach that disregards the ordering of words in

a document, leading to the loss of crucial contextual information that may affect the similarity scores and document ranking.

- Finally, the calculation of cosine similarity requires the use of a vector space model, which can be resource-intensive and computationally expensive, particularly for large document collections.

Applications:

- **Document retrieval:** Cosine similarity can be used to retrieve documents that are similar to a given query based on the similarity scores between the query and each document in the collection.

- **Text classification:** Cosine similarity can be used to classify documents into different categories based on their similarity to pre-defined category prototypes.

- **Collaborative filtering:** Cosine similarity can be used in recommendation systems to identify items that are similar to the user's preferences based on the similarity scores between the user's rating history and the item's feature vector.

- **Clustering:** Cosine similarity can be used to cluster similar documents together based on their similarity scores, allowing for efficient grouping and organization of large document collections.

- **Information extraction:** Cosine similarity can be used to identify key terms and entities in a document by comparing their similarity scores to the query or other documents in the collection.

- **Image retrieval:** Cosine similarity can be used to retrieve similar images based on their feature vectors, which can be represented as high-dimensional vectors in a vectorspace model.

- **Natural Language Processing:** Cosine similarity can be used to measure the similarity between word embeddings, which are high-dimensional vectors that represent words in a continuous vector space, allowing for efficient and accurate comparison of words and phrases.

Illustration of Cosine Similarity with example

Here are three documents. Let’s find out how similar are they?

d1- ant ant bee

d2 - dog bee dog hog dog ant dog

d3 - cat gnu dog eel fox

Step1 : Preparing Term-Text Table

Document Text Terms

Document	Text	Terms
d1	ant ant bee	ant bee
d2	dog bee dog hog dog ant dogcat gnu	ant bee dog hog cat dog
d3	dog eel fox	eel fox gnu

Step2 : Weighting by term frequency

Ex: length d1 = $\sqrt{(1^2+1^2)}$

	ant	bee	cat	Dog	eel	fox	gnu	hog	length
d1	2	1							$\sqrt{5}$
d2	1	1			4			1	$\sqrt{19}$
d3			1	1	1	1	1		$\sqrt{5}$

For ex: cos (θ) = (d1 .d2) / (|d1 | |d2 |) = (2.1 + 1.1 + 0.4 +0.1)/(√5 x √19) = 0.31

Final Cosine similarities between 3 documents are:

	d1	d2	d3
d1	1	0.31	0
d2	0.31	0	0.41
d3	0	0.41	1