Cosine Similarity

May 4, 2022

Definition: *Cosine Similarity* is the measure of similarity between two sequences of numbers. Mathematically this is shown as

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

for vectors $\mathbf{A}, \mathbf{B} \in R^n$.

Comment: Yes, we can use trigonometric measures to explain the similarity between two documents

Let's consider the following to statements:
1. "hello world"
2. "end world hunger"

If we look at the word counts for each of these we see the following table

Word	Statement 1 Count	Statement 2 Count
world	1	1
hello	1	0
end	0	0
hunger	0	0

We can then vectorize these. Let $\mathbf{A}$ and $\mathbf{B}$ represent statements 1 and 2 respectively. We observe

$$\mathbf{A} =< 1, 1, 0, 0 >;\ \mathbf{B} =< 1, 0, 1, 1 >$$

In this case, we consider vectors in $n = 4$ dimensional space.

Before we get into it, some definitions we need
Recall: Vector norm $||\mathbf{X}|| = \sqrt{\sum_{i=1}^{n} x_i^2}$ for $\mathbf{X} \in R^n$ (this represents a vectors length)
Recall: Dot product $\mathbf{X} \cdot \mathbf{Y} = \sum_{i=1}^{n} x_i y_i$ for $\mathbf{X}, \mathbf{Y} \in R^n$

So back in our example, using the definitions we can calculate the cosine similarity of these phrases as follows

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

$$= \frac{\sum_{i=1}^{4} x_i y_i}{\sqrt{\sum_{i=1} x_i^2}\sqrt{\sum_{i=1} x_i^2}} = \frac{(1)(1) + (1)(0) + (0)(1) + (0)(1)}{\sqrt{1^2 + 1^2 + 0^2 + 0^2}\sqrt{1^2 + 0^2 + 1^2 + 1^2}}$$

$$= \frac{1 + 0 + 0 + 0}{\sqrt{1+1}\sqrt{1+1+1}} = \frac{1}{\sqrt{2}\sqrt{3}} = 0.4082$$

So the cosine of the angle between these two vectors is 0.4082, which implies $\theta = 1.1503$ radians or 65.9082 degrees.

Intuitively, the closer the angle is to zero implies the more similar the vectors are. For instance, an angle of 0 implies the vectors are identical. Properties of cosine imply the closer this angle goes to 0, the closer the cosine of the angle is to 1. Therefore, we are looking to find cosine similarities closer to 1 for long positions and closer to 0 for short positions.