Cosine Similarity and Its Applications in AI

David Oniani
Luther College
oniada01@luther.edu

April 19, 2020

Abstract

Choosing the right metric [1] can be crucial to designing performant artificial intelligence models. Thousands of packages and libraries have been built and written just for providing these metrics. Cosine similarity is one of many metrics used extensively in natural language processing and artificial intelligence tasks. The paper will introduce the technique and discuss its advantages and disadvantages as well as compare it to other approaches. Additionally, sample implementations of the above-mentioned approaches will also be provided.

Table of Contents

1 Introduction

There are a number of approaches for comparing whether two texts are semantically similar to each other. Cosine similarity is one of those methods. In order to understand how cosine similarity works, let us first discuss the modeling of a text document and cosine similarity in general and then proceed by its applications in text document similarity tasks.

There are several ways in which a text document can be modeled. This includes a bag of words modeling, where the frequency of a term in a text document represents its weight and therefore, more frequent words are going to be deemed more "important." The whole idea behind a text document modeling is to quantify the textual data into the numeric data (usually, vectors). Once the numeric data is obtained, we can then apply various text semantic similarity techniques in order to compare whether two documents are similar to each other.

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them [2]. If we have two vectors $\vec{a}$ and $\vec{b}$, then the cosine of these two vectors is $\vec{a} \cdot \vec{b}$ which is equal to $\|\vec{a}\| \left\|\vec{b}\right\| \cos\theta$ where θ is the angle between these vectors (Euclidean dot product formula). It also can be represented as the product of two vectors. Therefore if $\vec{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and $\vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$, then $\vec{a} \cdot \vec{b}$ is also equal to $a_1b_1 + a_2b_2 + a_3b_3$.

Suppose that we have two documents D_1 and D_2 modeled as term vectors $\vec{t_1}$ and $\vec{t_2}$ respectively. Then the similarity of two documents corresponds to the correlation between the vectors and can be quantified as a cosine of the angle between the vectors. The formula would be

$$SIM(D_1, D_2) = \frac{\vec{t_1} \cdot \vec{t_2}}{|\vec{t_a}| \times |\vec{t_b}|}$$

As a result, the cosine similarity is non-negative, bounded by the closed interval $[0, 1]$.

As an example, consider two vectors $\vec{a} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ and $\vec{b} = \begin{bmatrix} 5 \\ 12 \end{bmatrix}$. Then the size of a is $\sqrt{1^2 + 2^4} = 5$ and the size of b is $\sqrt{3^2 + 4^2} = 13$. Then $A\dot{B} = 15 + 48 = 63$ and $|\vec{a}| \times |\vec{b}| = 5 * 13 = 65$. Therefore the cosine of an angle between these vectors is $\frac{63}{65}$ which is the similarity measure between these two vectors.

References

[1] Rachel Thomas and David Uminsky. "The Problem with Metrics is a Fundamental Problem for AI". In: (2020). URL: https://arxiv.org/abs/2002.08512.

[2] Wikipedia The Free Encyclopedia. *Cosine similarity*. URL: https://en.wikipedia.org/wiki/Cosine_similarity.