Similarity measures

How to compare vectors

When comparing vectors:

- Two vectors that exactly the same will have a relation of 1.
- Two vectors that are exactly different will have a relation of 0.

There is an exception when using cosine due to its nature:

- When using *non-negative* numbers the range is between 0 and 1
- When using *negative* numbers the range is between -1 and 1

Euclidean Distance and Similarity

$E(A,B) = \sqrt{(A_1-B_1)^2 + (A_2-B_2)^2 + (A_n-B_n)^2}$ --This is the Euclidean Distance formula

$1 / (1 + E(A,B))$ -- This is the Euclidean Similarity formula

Example 1.

A = [0, 1]; B = [1, 1]

$E(A,B) = \sqrt{(A_1-B_1)^2 + (A_2-B_2)^2} = \sqrt{(0-1)^2 + (1-1)^2} = 1$; the distance between the two points is 1.

$1 / (1 + E(A,B)) = 1 / (1 + 1) = .5$; The similarity between the two points is .5

Euclidean Distance and Similarity

Further examples:

Distance | Similarity

A = [0, 1, 1, 1]; $D(A,B) = \sqrt{(0-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2} = 1$ | .500

B = [1, 1, 1, 1]; $D(B,C) = \sqrt{(1-1)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2} = \sqrt{2}$ | .414

C = [1, 1, 0, 0]; $D(B,D) = \sqrt{(0-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2} = \sqrt{3}$ | .366

D = [1, 0, 0, 0]; $D(C,D) = \sqrt{(0-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2} = 1$ | .500

Euclidean Distance and Similarity

“Euclidean distance measures the straight line distance between two points in n-dimensional space.

Use-cases for the Euclidean Distance algorithm

We can use the Euclidean Distance algorithm to work out the similarity between two things. We might then use the computed similarity as part of a recommendation query. For example, to get movie recommendations based on the preferences of users who have given similar ratings to other movies that you’ve seen.”

https://neo4j.com/docs/graph-algorithms/current/algorithms/similarity-euclidean/

Jaccard Similarity (Coefficient)

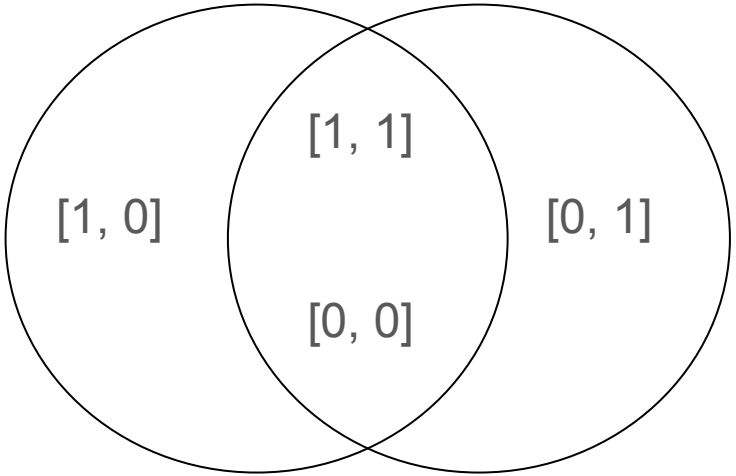
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad \text{or} \quad J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

| A |, | B | = Total items in set A, or B; | A ∩ B | = Number of items in common

Example 1.

A = [0, 1]; B = [1, 1]

$$J(A,B) = \frac{1}{2 + 2 - 1} = \frac{1}{3}$$



Jaccard Similarity (Coefficient)

Further examples:

A = [0, 1, 1, 1]; J(A,B) = 3 / (4 + 4 - 3) = 3 / 5 = .6

B = [1, 1, 1, 1]; J(B,C) = 2 / (4 + 4 - 2) = 2 / 6 = .33~

C = [1, 1, 0, 0]; J(B,D) = 1 / (4 + 4 - 1) = 1 / 7 = .143

D = [1, 0, 0, 0]; J(C,D) = 3 / (4 + 4 - 3) = 3 / 5 = .6

Jaccard Similarity (Coefficient)

“Jaccard similarity (coefficient), a term coined by Paul Jaccard, measures similarities between sets. It is defined as the size of the intersection divided by the size of the union of two sets.

Use-cases for the Jaccard Similarity algorithm

We can use the Jaccard Similarity algorithm to work out the similarity between two things. We might then use the computed similarity as part of a recommendation query. For example, you can use the Jaccard Similarity algorithm to show the products that were purchased by similar customers, in terms of previous products purchased.”

https://neo4j.com/docs/graph-algorithms/current/algorithms/similarity-jaccard/

Cosine Similarity

$$C(A,B) = \frac{A \cdot B}{|A||B|} \quad \text{alternatively} \quad \begin{matrix} A = [X_1, Y_1] \\ B = [X_2, Y_2] \end{matrix} \quad = \quad \frac{X_1 * X_2 + Y_1 * Y_2}{\sqrt{(X_1^2+Y_1^2) * (X_2^2+Y_2^2)}}$$

Example 1.

$$\begin{matrix} A = [0, 1] \\ B = [1, 1] \end{matrix} \quad = \quad \frac{0 * 1 + 1 * 1}{\sqrt{(0^2+1^2) * (1^2+1^2)}} \quad = \quad \frac{1}{\sqrt{2}} \quad = \quad .7071$$

Cosine Similarity

Further examples:

A = [0, 1, 1, 1]; C(A,B) = $0*1+1*1+1*1+1*1 / \sqrt{(0^2+1^2+1^2+1^2)(1^2+1^2+1^2+1^2)} = .8660$

B = [1, 1, 1, 1]; C(B,C) = $1*1+1*1+1*0+1*0 / \sqrt{(1^2+1^2+1^2+1^2)(1^2+1^2+0^2+0^2)} = .7071$

C = [1, 1, 0, 0]; C(B,D) = $1*1+1*0+1*0+1*0 / \sqrt{(1^2+1^2+1^2+1^2)(1^2+0^2+0^2+0^2)} = .5000$

D = [1, 0, 0, 0]; C(C,D) = $1*1+1*0+0*0+0*0 / \sqrt{(1^2+1^2+0^2+0^2)(1^2+0^2+0^2+0^2)} = .7071$

Cosine Similarity

“The cosine similarity is the cosine of the angle between two *n*-dimensional vectors in an *n*-dimensional space. It is the dot product of the two vectors divided by the product of the two vectors' lengths (or magnitudes).

Use-cases for the Cosine Similarity algorithm

We can use the Cosine Similarity algorithm to work out the similarity between two things. We might then use the computed similarity as part of a recommendation query. For example, to get movie recommendations based on the preferences of users who have given similar ratings to other movies that you’ve seen.”

https://neo4j.com/docs/graph-algorithms/current/algorithms/similarity-cosine/

Recap of similarity scores

Data	Set	Euclidean	Jaccard	Cosine
A = [0, 1, 1, 1]	(A, B)	.5000	.6000	.8660
B = [1, 1, 1, 1]	(B, C)	.4142	.3333	.7071
C = [1, 1, 0, 0]	(B, D)	.3660	.1429	.5000
D = [1, 0, 0, 0]	(C, D)	.5000	.6000	.7071

So that's how binary data works, what about text?

A = She sells seashells on the seashell shore.

B = The seashells she sells are seashore shells

A ∪ B	she	sell	seashell	on	the	shore	are	seashore	shell
A	1	1	2	1	1	1	0	0	0
B	1	1	1	0	1	0	1	1	1
J(A, B)	1	1	.5	0	1	0	0	0	0

Build a list of all the words that are going to be compared. Assign values based on the occurrence of the words in the text. Calculate their similarity!

J(A, B) = 3.5 / (7 + 7 - 3.5) = .333~; **C(A,B)** = 0.6299; **E(A, B)** = 0.2898