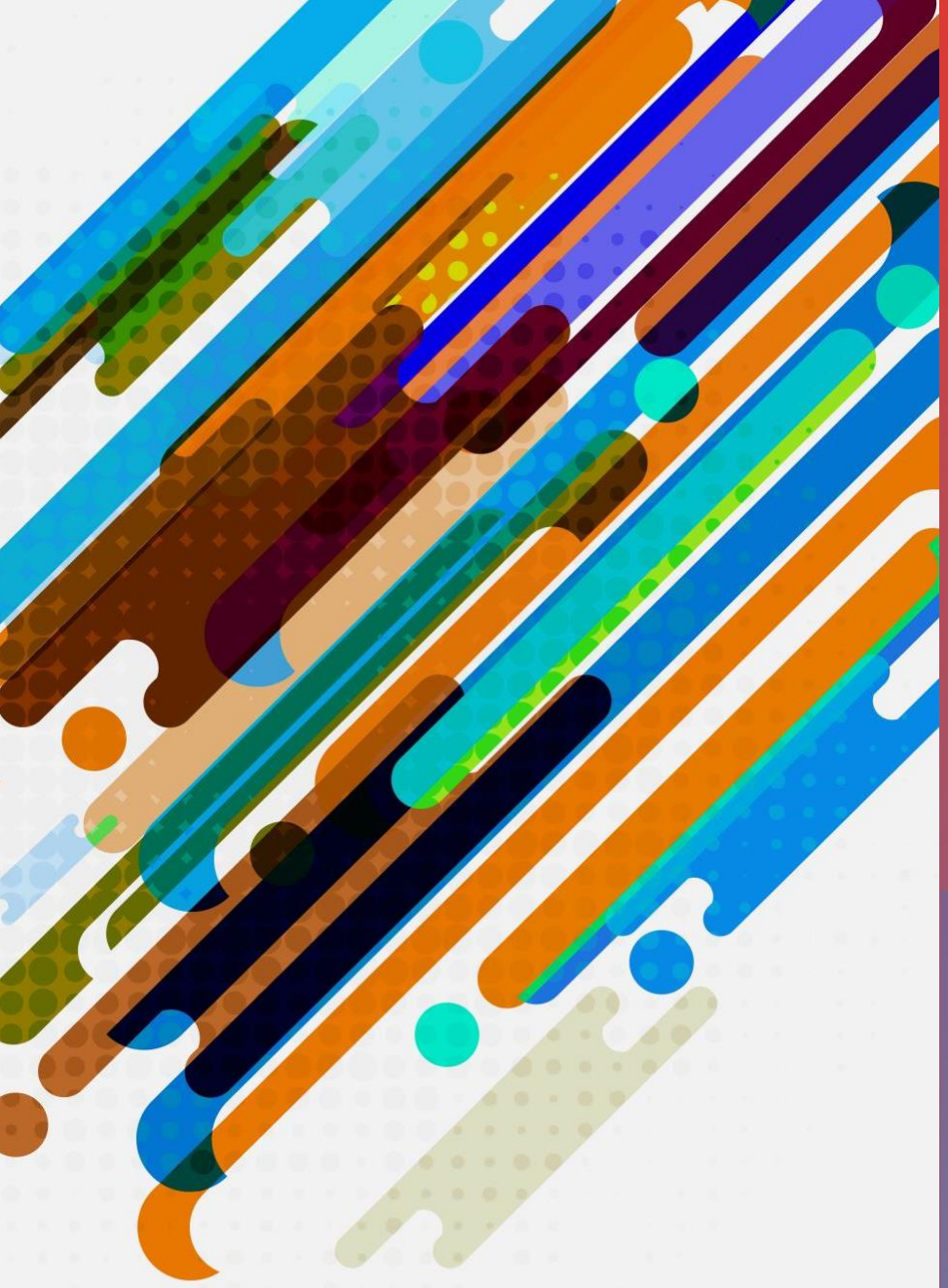


TEXT SIMILARITY

NLP

Echemos un vistazo a estos pares de oraciones, ¿cuál de estos pares creen que tienen oraciones similares?

Sorprendentemente, lo contrario es cierto para los modelos de PNL. De acuerdo con la forma en que funciona la similitud de texto en NLP, las oraciones de los dos últimos pares son muy similares, ¡pero no las de los dos primeros!

¿CÓMO SE CALCULA LA SIMILITUD?

Oración 1: “Global warming is here”

Oración 2: “Ocean temperature is rising”

¿CÓMO SE CALCULA LA SIMILITUD?

PASO 1

Escoge solo las palabras únicas de las dos oraciones, lo que equivaldría a *7*.

Palabras únicas: *global, warming, is, here, ocean, temperature, rising*

¿CÓMO SE CALCULA LA SIMILITUD?

PASO 2

Contar el número de apariciones de palabras únicas en cada una de las oraciones

***Análisis de la oración 1*:**

```
global, 1
warming, 1
is, 1
here, 1
ocean, 0
temperature, 0
rising, 0
```

***Análisis de la oración 2*:**

```
global, 0
warming, 0
is, 1
here, 0
ocean, 1
temperature, 1
rising, 1
```

¿CÓMO SE CALCULA LA SIMILITUD?

PASO 3

Convertir a vectores el número de apariciones de palabras únicas en cada una de las oraciones

***Análisis de la oración 1*:**

```
global, 1
warming, 1
is, 1
here, 1
ocean, 0
temperature, 0
rising, 0
```

[1, 1, 1, 1, 0, 0, 0]

***Análisis de la oración 2*:**

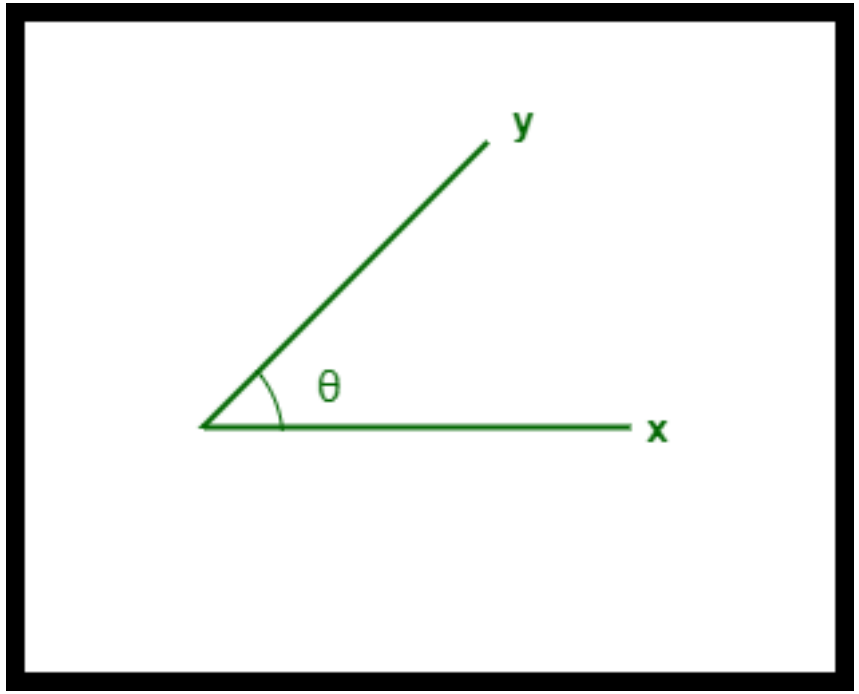
```
global, 0
warming, 0
is, 1
here, 0
ocean, 1
temperature, 1
rising, 1
```

[0, 0, 1, 0, 1, 1, 1]

SIMILITUD COSENO

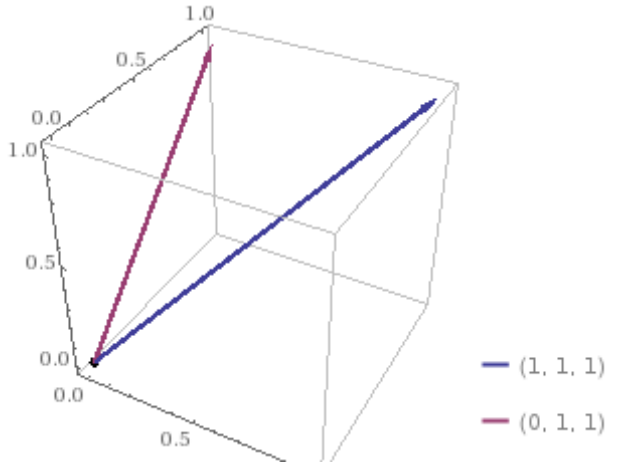
La similitud coseno es una medida de la similitud existente entre dos vectores en un espacio que posee un producto interior con el que se evalúa el valor del coseno del ángulo comprendido entre ellos. Esta función trigonométrica proporciona un valor igual a 1 si el ángulo comprendido es cero, es decir si ambos vectores apuntan a un mismo lugar. Cualquier ángulo existente entre los vectores, el coseno arrojaría un valor inferior a uno.

En minería de textos se aplica la similitud coseno con el objeto de establecer una métrica de semejanza entre textos.

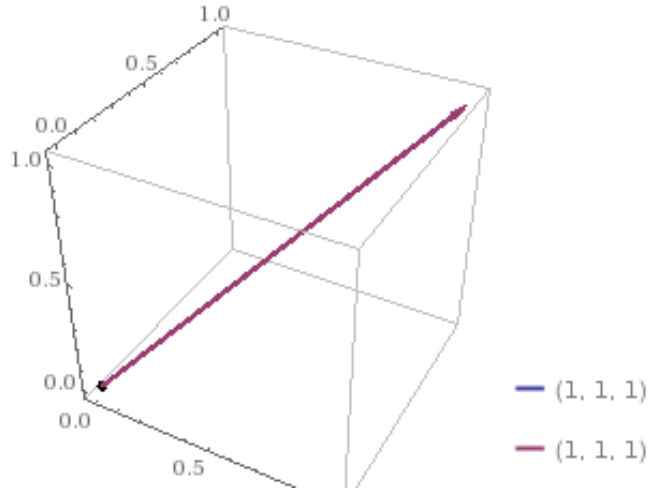


SIMILITUD COSENO

Solo nos interesa el ángulo entre estos dos vectores. Cuanto más cerca estén las dos líneas, menor será el ángulo y, por lo tanto, aumentará la similitud. Por lo tanto, si dos oraciones son perfectamente similares, solo vería una línea en el espacio 3D, ya que las dos líneas se superpondrían entre sí.



A medida que los vectores se acercan, la similitud aumenta, ya que ambas oraciones tienen 2 palabras en común.



¡Una combinación perfecta de 2 oraciones!

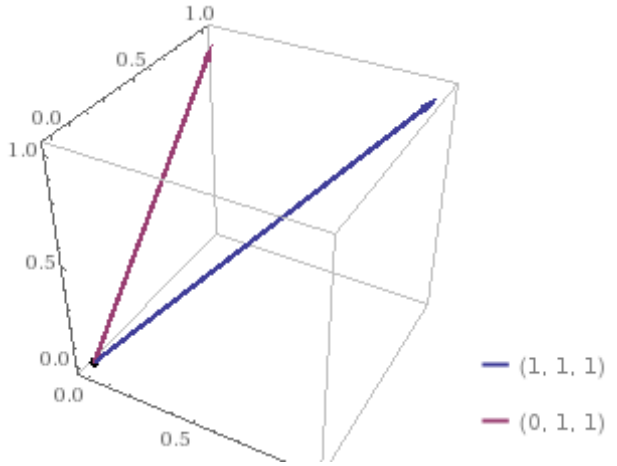
SIMILITUD COSENO

Para calcular la similitud del coseno, usaremos esta fórmula:

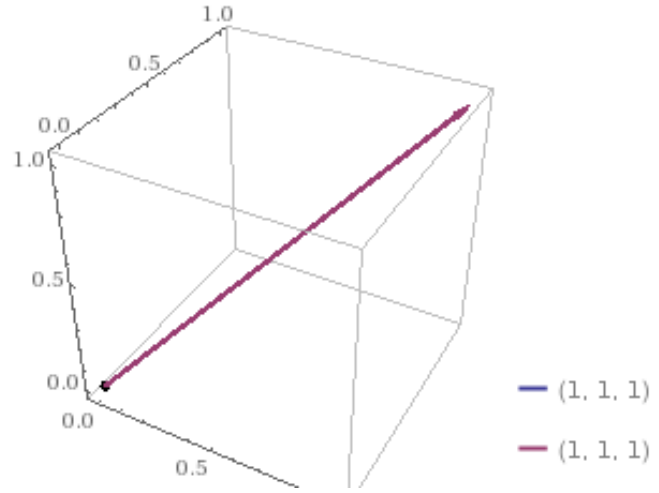
$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Here's a breakdown of the formula:

- $\cos(\theta)$ is the cosine of the angle θ between the two vectors.
- $A \cdot B$ is the dot product of vectors A and B.
- $\|A\|$ and $\|B\|$ are the magnitudes (or norms) of vectors A and B, respectively.
- $\sum_{i=1}^{n} A_i B_i$ is the sum of the products of the corresponding entries of the vectors.
- $\sum_{i=1}^{n} A_i^2$ and $\sum_{i=1}^{n} B_i^2$ are the sums of the squares of the components of vectors A and B, respectively.



A medida que los vectores se acercan, la similitud aumenta, ya que ambas oraciones tienen 2 palabras en común.



¡Una combinación perfecta de 2 oraciones!

¿CÓMO SE CALCULA LA SIMILITUD?

PASO 4

Calcular el ángulo del coseno entre los dos vectores.

1. Averigüemos el producto punto para nuestro caso:

(1*0) + (1*0) +(1*1) +(1*0) +(1*0) +(1*0) +(1*0) = 1

2. Calcular el tamaño (la norma) de los dos vectores:

Length of vector 1 -> $\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2}$

Length of vector 2 -> $\sqrt{0^2 + 0^2 + 1^2 + 0^2 + 1^2 + 1^2 + 1^2}$

El resultado es $\frac{1}{4} = 0.25$. Por tanto, las dos oraciones son solo un 25% similares, lo que es completamente opuesto a lo que mostraría el análisis semántico.