Sprint 2 - Matemáticas
Interacción Sociotecnológica
Daniel Gómez

INTRODUCCIÓN

El objetivo de este sprint es familiarizarnos con los conceptos matemáticos básicos para implementar un algoritmo de recomendación. En el sprint anterior hemos logrado cargar una base de datos desde un formato csv y ordenar sus elementos de acuerdo a las filas. En este sprint sentaremos las bases matemáticas necesarias para hacer un análisis de similitud entre los elementos.

Recordemos que una base de datos como la que hemos usado puede verse así:

nombre	característica 1	característica 2	característica 3	característica 4
ana	99	63	65	62
juan	36	28	87	86
tatiana	18	73	63	41
oscar	68	89	2	23

La idea es definir qué tan parecido es un elemento con respecto a otro. Por ejemplo: qué tanto se parece ana a juan. Para esto tenemos que calcular la similitud entre ana y juan (o entre cualquier pareja). Existen muchísimos tipos de algoritmos para calcular la similitud, pero para este ejercicio exploraremos la similitud coseno. A continuación presentamos paso a paso cómo se calcula.

1. Producto punto

El primer paso es recordar qué es el producto punto. SI tenemos dos vectores A=[1,3,5] y B=[2,-1,4] el producto punto entre estos dos vectores se escribe como A·B y es equivalente a: multiplicar el elemento n del vector A con el elemento n del vector B y luego de multiplicar cada uno, hacemos la suma de todas las multiplicaciones. Ejemplo

A·B = (1 * 2) + (3 * -1) + (5 * 4) = 2 - 3 + 20 = 19

En nuestra base de datos el producto punto podría hacerse entre cualquiera de las personas que hacen parte de ella, por ejemplo entre tatiana y oscar.

2. Magnitud

El segundo paso consiste en repasar la idea del magnitud de un vector. SI tenemos un vector A=[1,3,5] su magnitud se describe como ||A|| y equivale a:

$$\sqrt{1^2 + 3^2 + 5^2} = \sqrt{1+9+25} = \sqrt{35} = 5.916079783099616$$

3. La similitud coseno

Finalmente, la similitud coseno se mide basada en las dos ideas anteriores y nos permite definir qué tan parecido es un vector con otro. La similitud coseno entre dos vectores es el producto punto dividido por la magnitud de cada uno.

$$\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

si tenemos dos vectores A=[1,3,5] y B=[2,-1,4] la similitud coseno entre ellos la calculamos haciendo tres operaciones entre los vectores y luego haciendo una división.

Primero calculamos A·B
A·B = (1 * 2) + (3 * -1) + (5 * 4) = 2 - 3 + 20 = 19

Luego calculamos ||A|| y ||B||
||A|| = $\sqrt{1^2 + 3^2 + 5^2} = \sqrt{1+9+25} = \sqrt{35} = 5.916079783099616$
||B|| = $\sqrt{2^2 + (-1)^2 + 4^2} = \sqrt{4+1+16} = \sqrt{21} = 4.58257569495584$

y finalmente hacemos la división:

similitud coseno = 19/($5.916079783099616 * 4.58257569495584$)
= 19 / 27.11048
= 0.7008

EJERCICIO

Al finalizar el sprint debemos tener un programa que permita cargar una base de datos en formato csv que tenga por lo menos 5 características numéricas con valores entre 0 y 1. Luego, calcular la similitud coseno entre dos de los elementos de la base de datos que seleccionemos de manera interactiva desde una lista desplegable. La interfaz debe tener dos listas desplegables, un botón de “ejecutar” y finalmente un espacio en el que se presente el resultado de la distancia. Debo poder seleccionar dos personas cualquiera seleccionando sus nombres en las listas desplegables y calcular la similitud coseno entre ellas. El sprint lo evaluamos en la clase siguiente.