VECTOR SPACE MODEL.

Nos dan una base de datos con documentos. 5 documentos.

d1 → Information Retrieval Systems
d2 → Information Storage
d3 → Digital Speech Synthesis Systems.
d4 → Speech Filtering
d5 → Speech Retrieval.

Nos dan la consulta (Query) que queremos recupear:

Query → Inform. Speech filtering, (1 query) Speech Retrieval, (1 query) (nos dan 2).

1) Tabla de frecuencias (matriz de frecuencias).

	Digital	filtering	Inform.	Retrieval	Speech	Storage	Synthesis	Systems
d1	0	0	1	1	0	0	0	1
d2	0	0	1	0	0	1	0	0
d3	1	0	0	0	1	0	1	1
d4	0	1	0	0	1	0	0	0
d5	0	0	0	1	1	0	0	0
sum	1	1	2	2	3	1	1	2

- Hacemos la tabla y vamos rellenando donde tengamos la palabra y completamos la suma de todo también.

2) Frecuencia inversa de documentos

$$idf = \log\left(\frac{n^{\circ}\text{ documentos totales}}{\text{suma total de la palabra}}\right)^{5}$$ → aplicamos la fórmula a cada palabra

Palabra	DOC-frecuency	IDF
Digital	1	log (5/1) = 0,699
Filtering	1	log (5/1) = 0,699
Information	2	log (5/2) = 0,398
Retrieval	2	log (5/2) = 0,398
Speech	3	log (5/3) = 0,222
Storage	1	log (5/1) = 0,699
Synthesis	1	log (5/1) = 0,699
Systems	2	log (5/2) = 0,398

3. Calcular la matriz tf.idf.

- multiplicamos la frecuencia de la palabra por su idf que hemos sacado y la longitud del vector.

$$Lenght = \sqrt{palabra1^2 + palabra2^2 + palabra3^2 \ldots}$$

↳ de cada diccionario hacemos un Leght.
↳ ponemos los palabras que aparezcan en el diccionario, (su idf)

	Digital	Filtering	Information	Retrieval	Speech	Storage	Syn	Systems
d1	0	0	1x0,397	1x0,398	0	0	0	1x0,398
d2	0	0	1x0,397	0	0	1x 0,699	0	0
d3	1x0,699	0	0	0	1x0,222	0	1x0,699	1x 0,398
d4	0	1x0,699	0	0	1x0,222	0	0	0
d5	0	0	0	1x0,398	1x0,222	0	0	0

Lenght $d1 = \sqrt{0,398^2 + 0,398^2 + 0,398^2} = 0,689$.

Leight $d2 = 0,804$ // $Ld3 = 1,088$ // $d4 = 0,733$ // $d5 = 0,456$.

4) Vector de consulta y consulta

Consulta/Query es: "Information (1), Speech, Filtering", "Speech (2) Retrieval"

↳ Frecuencia máxima de un término — Speech → (2) veces

Vector Query = $\frac{\text{Frecuencia del término}}{\text{Frec. máx de los term. de la Query}}$ x idf del término.

	Digital	filtering	Information	Retriev.	Speech	Storage	Synthe	Syst	Leight
q	$\frac{1}{2}$ x 0,699	$\frac{1}{2}$ x 0,699 = 0,349	$\frac{1}{2}$ x 0,398 = 0,199	$\frac{1}{2}$ x 0,398 = 0,199	$\frac{2}{2}$ x 0,222 = 0,222	$\frac{1}{2}$ x 0,699	$\frac{1}{2}$ x 0,699	$\frac{1}{2}$ x 0,398	0,501.

(!!) Lenght = $\sqrt{\text{de cada resultado}^2 + \ldots}$

(!!) De esos no se hace porque no sale a la query.

5) Función de similitud.

- Multiplicar el vector de la query x vector de cada documento / multipl. de longitudes.

Función de similitud = $\cos \frac{\text{vector (4)} \times \text{vector (3)}}{\text{Leight (4)} \times \text{Leight (3)}}$

sólo las palabras que salgan

$\cos Sim(d1, q) = \frac{(0,397 \times 0,199) + (0,398 \times 0,199)}{0,501 \times 0,689} = [0,459]$ (Informat, Retrieval)

$\cos Sim(d2, q) = \frac{(0,397 \times 0,199)}{0,501 \times 0,804} = [0,196]$

$\cos Sim(d3, q) = \frac{0,222 \times 0,222}{0,501 \times 1,088} = [0,090]$

$\cos Sim(d4, q) = \frac{(0,222 \times 0,222) + (0,699 \times 0,349)}{0,501 \times 0,733} = [0,799]$

$\cos Sim(d5, q) = \frac{(0,398 \times 0,199) + (0,222 \times 0,222)}{0,501 \times 0,456} = [0,562]$

(!!) Cuanto más grande sea el cosSim, más similar será el doc a la query.

Por lo tanto el orden de mayor similitud a menor será:

$[d4 \rightarrow d5 \rightarrow d1 \rightarrow d2 \rightarrow d3]$