

MODELO VECTORIAL: SIMILITUD COSENO

Considérese una colección de 1.000 documentos entre los cuales se encuentran los siguientes:

Doc1: **programmers** build **computer software**

Doc2: most **software** has **bugs**, but good **software** has less **bugs** than bad **software**

Doc3: some **bugs** can be found only by executing the **software**, not by examining the source **code**

Los términos a considerar se han indicado en negrita.

Se pide calcular la similitud coseno entre la consulta “**computer software programmers**” y cada uno de los documentos (esquema de pesado Inc.Itc). En la tabla se indica el *df* de cada término considerado. Se han calculado los resultados truncando a dos decimales.

DEFINICIONES:

$$tf_{t,d} = \begin{cases} 1 + \log_{10} f_{t,d}, & \text{si } f_{t,d} > 0 \\ 0, & \text{otro caso} \end{cases}$$

$$idf_t = \log_{10} (N/df_t)$$

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} = \frac{\vec{q}}{|\vec{q}|} \cdot \frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|k|} q_i d_i}{\sqrt{\sum_{i=1}^{|k|} q_i^2} \sqrt{\sum_{i=1}^{|k|} d_i^2}}$$

Term			Consulta				Doc1				Doc2				Doc3			
	df _t	idf _t	f _{t,q}	tf _{t,q}	W _{t,q} =tf _{t,q} idf _t	L-Norm	f _{t,d}	tf _{t,d}	W _{t,d} =tf _{t,d} idf _t	L-Norm	f _{t,d}	tf _{t,d}	W _{t,d} =tf _{t,d} idf _t	L-Norm	f _{t,d}	tf _{t,d}	W _{t,d} =tf _{t,d} idf _t	L-Norm
bugs	50	1,3	0	0	0	0,00	0	0	0	0,00	2	1,3	1,3	0,66	1	1	1	0,57
code	20	1,69	0	0	0	0,00	0	0	0	0,00	0	0	0	0,00	1	1	1	0,57
computer	100	1	1	1	1	0,45	1	1	1	0,57	0	0	0	0,00	0	0	0	0,00
programmers	20	1,69	1	1	1,69	0,76	1	1	1	0,57	0	0	0	0,00	0	0	0	0,00
software	100	1	1	1	1	0,45	1	1	1	0,57	3	1,47	1,47	0,74	1	1	1	0,57

Esquema de pesado Inc.Itc:

- para los **documentos** log-pesado, no idf y normalización coseno;
- para la **consulta** log-pesado, idf y normalización coseno.
-

Similitud coseno(consulta,Doc1)= 0.93= 0+0+(0.45x0.57)+(0.76x0.57)+(0.45x0.57)

Similitud coseno(consulta,Doc2)= 0.33=0+0+0+0+(0.45x0.74)

Similitud coseno(consulta,Doc3)= 0.25=0+0+0+0+(0.45x0.57)