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 lnc.ltc). En la tabla se indica el *df* de cada término considerado. Se han calculado los resultados truncando a dos decimales.

DEFINICIONES:

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

$$\text{idf}_t = \log_{10}(N/\text{df}_t)$$

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

Term			Consulta				Doc1				Doc2				Doc3			
	df_t	idf_t	$f_{t,q}$	tf_{tq}	$W_{t,q}$=tfxidf	L-Norm	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tfxidf	L-Norm	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tfxidf	L-Norm	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tfxidf	L-Norm
bugs	50																	
code	20																	
computer	100																	
programmers	20																	
software	100																	

Esquema de pesado lnc.ltc:
- para los **documentos** log-pesado, no idf y normalización coseno;
- para la **consulta** log-pesado, idf y normalización coseno.