Hoja de problemas

En una colección de 806791 documentos, disponemos de la siguiente información sobre unos términos y unos documentos en particular:

term	document frequency	Doc1	Doc2	Doc3
car	18165	27	4	24
auto	6723	3	33	0
insurance	19,241	0	39	29
best	25235	14	9	17

1 Calcula el valor tf-idf para estos términos y documentos.

2 ¿Cuál es la similitud coseno entre "best car best insurance" y los documentos Doc 1, Doc 2 y Doc 3, respectivamente? Usa el esquema ltn.lnc.

3 ¿Cuál de los tres documentos tendría la mejor puntuación utilizando el esquema ltn.lnc?

Solución Step 1 (1.2 points) : Compute tf-idf values

term	Doc1	Doc2	Doc3
car	4.0057	2.6394	3.9215
auto	3.0712	5.2365	0.0000
insurance	0.0000	4.2041	3.9953
best	3.2294	2.9407	3.3563

Step2 (0.4 points): Compute vector representation for query

term	query
car	1.0
auto	0.0
insurance	1.0
best	1.301

Step 3 (1.2 points): The cosine similarity between the query and documents 1, 2 and 3 is:

$cosine(query, Doc_1) = \frac{1*4.0057+1*0+1.301*3.2294}{\sqrt{1^2+0^2+1^2+1.301^2}} = \frac{8.2071}{1.9216} = 4.27$
$cosine(query, Doc_2) = \frac{1*2.6394+1*4.2041+1.301*2.9407}{\sqrt{1^2+0^2+1^2+1.301^2}} = \frac{10.6694}{1.9216} = 5.55$
$cosine(query, Doc_3) = \frac{1*3.9215+1*3.9953+1.301*3.3563}{\sqrt{1^2+0^2+1^2+1.301^2}} = \frac{12.2833}{1.9216} = 6.39$

Step 4 (0.2 points): The document of highest rank is Doc3