Vector space exercise

Tenemos una base de datos de documentos de 5 documentos con el siguiente contenido
Document 1(**d1**):"Information Retrieval Systems"
Document 2(**d2**):"Information Storage"
Document 3(**d3**):"Digital Speech Synthesis Systems"
Document 4(**d4**):"Speech Filtering"
Document 5(**d5**):"Speech Retrieval"

Queremos recuperar los documentos de esa base de datos que coincidan mejor con mis necesidades de información. Para eso, la consulta es**: Information Speech Filtering, Speech Retrieval.**

Pasos

1. ***Matriz de frecuencia***
2. ***Frecuencia inversa del documento:***
3. ***Vector de consulta***
4. ***Función de similitud***

1.-Matriz de frecuencia: calcular la frecuencia de cada término en cada documento

	Digital	Filtering	Information	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

2.-Frecuencia inversa de documentos (número de documentos / frecuencia de los términos en todos los documentos). IDF es una estadística numérica que está destinada a reflejar lo importante que es una palabra para un documento en una colección o corpus. ayuda a ajustarse al hecho de que algunas palabras aparecen con más frecuencia en general

TERM	DOC-FREQUENCY	IDF
Digital	1	log(5/1)=0.699
Filtering	1	log(5/1)=0.699
Information	2	**log(5/2)= 0.397**
Retrieval	2	log(5/2)=0.398
Speech	3	**log(5/3)= 0.221**
Storage	1	log(5/1)=0.699
Synthesis	1	log(5/1)=0.699
Systems	2	**log(5/2)= 0.397**

3.- Calcular la matriz tf.idf - multiplicar la frecuencia x el IDF del término y la longitud del vector (última columna)
Length of d1=sqrt(_____^2+_____^2+_____^2)=
Length of d2=sqrt(_____^2+_____^2)=
Length of d3=sqrt(0.699^2+0.222^2+0.699^2+0.398^2)=1.088
Length of d4=sqrt(0.699^2+0.222^2)=0.733
Length of d5=sqrt(0.398^2+0.222^2)=0.456

	Digital	Filtering	Information		Retrieval	Speech	Storage	Synthesis	Systems	**Length**
d1	0	0	0'397		0'398	0	0	0	0'397	
d2	0	0	0'397		0	0	0'6641	0	0	
d3	1x0.699	0	0		0	1x0.222	0	1x0.699	0.398	**1.088**
d4	0	0.699	0		0	0.222	0	0	0	**0.733**
d5	0	0	0		0.398	0.222	0	0	0	**0.456**

4.-Vector de consulta y consulta
The Query is: *Information Speech Filtering, Speech Retrieval*
La frecuencia máxima de un término es ("Speech")=2
Query vector: frecuencia del término/frecuencia máxima de los términos de la consulta) X idf del término en la base de datos de documentos.

Q[0 ; (1/2)*0.699=**0.349;** (1/2)*0.398=**0.199;** (1/2)*0.398=**0.199;** (_____)*0.222=**_____;**
0 ; 0; 0]

Length= sqrt(_____^2+_____^2+_____^2+_____^2)=**_____**

	Digital	Filtering	Information	Retrieval	Speech	Storage	Synthesis	Systems	**Length**
d1		(1/2)*0.699=0.349							**0,501**

7.-Función de similitud: multiplicar el vector de la consulta por el vector de cada documento dividido por la multiplicación de sus longitudes

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

cosSim(d1,q)=(_____*_____+_____*_____)/(_____*_____)=_____
cosSim(d2,q)=(0.398*0.199)/(0.501*0.804)=**0.197**
cosSim(d3,q)=(0.222*0.222)/(0.501*1.088)=**0.090**
cosSim(d4,q)=(0.222*0.222+0.699*0.349)/(0.501*0.733)=**0.799**
cosSim(d5,q)=(_____*_____+_____*_____)/(0.501*0.456)=_____
cuanto más grande el coseno más similar será le documento a la consulta. Así, el orden de presentación es:

_____