

Ranking Retrieval

Profesor Heider Sanchez

P1. Normalización de la longitud:

Se tiene la siguiente tabla de conteos para la consulta y dos documentos.

<i>i</i>	term	Frequency (<i>tf</i>)		
		Q	Doc1	Doc2
1	afecto	115	15	30
2	celosa	10	5	0
3	chisme	2	20	22
4	borrascoso	0	25	0

Se le pide aplicar las siguientes técnicas de scoring:

- a) El score es la sumatoria del *log-frequency weighth* de los términos comunes.

$$Score1(Q, D) = \sum_{i \in Q \cap D} q_i \cdot d_i$$

En donde, $q_i = \log_{10}(1 + tf_{i,Q})$ y $d_i = \log_{10}(1 + tf_{i,D})$

- b) El score ahora se aplica con los vectores unitarios.

$$Score2(Q, D) = \sum_{i \in Q \cap D} nq_i \cdot nd_i$$

En donde, $nq_i = q_i / \|\vec{Q}\|_2$, de igual forma para nd_i .

$$\|\vec{Q}\|_2 = \sqrt{\sum_{i=1}^{|Q|} q_i^2}$$

Llene el siguiente cuadro, analice los resultados y de una explicación de dicho comportamiento.

	(Q, Doc1)	(Q, Doc2)
Score1	3.93	3.73
Score2	0.69	0.78
Explicación		

P2. TF-IDF:

Dada la siguiente tabla en donde se distribuye los pesos TF-IDF para dos documentos de la colección y para la consulta, se pide calcular el score sin normalizar (dot-product) y el score normalizado (cosine similarity) entre Q y cada documento.

Id	Término	Doc1 (TF-IDF)	Doc2 (TF-IDF)	Q (TF-IDF)
T1	Clima	1.452	0	0
T2	Biblioteca	0	2.093	1.345
T3	Universidad	2.122	0	1.453
T4	Alcalá	3.564	0	1.987
T5	España	4.123	4.245	0
T6	Libros	0	1.234	2.133
T7	Geografía	0	0	0
T8	Población	2.342	0	0
T9	Electricidad	0	0	0
T10	Ciencia	0	0	0
T11	Social	0	2.345	0
T12	Luz	1.975	0	0
T13	Unamuno	4.543	2.135	3.452
T14	Física	0	0	0
T15	Fluidos	6.134	0	0
T16	Literatura	2.234	3.456	4.234

$$DotProduct(Q, D) = \sum_{i \in Q \cap D} q_i \cdot d_i$$

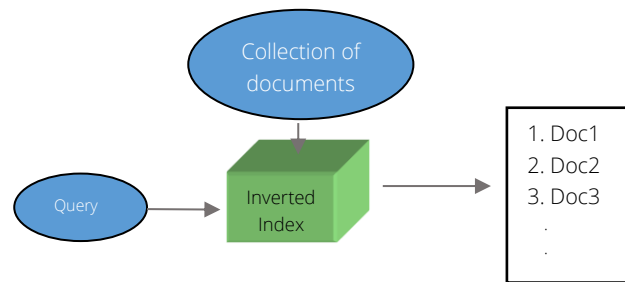
$$cosine(Q, D) = \frac{\sum_{i \in Q \cap D} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{|Q|} q_i^2 \times \sum_{i=1}^{|D|} d_i^2}}$$

En donde q_i es el peso tf-idf del término T_i respecto al documento Q. Lo mismo con d_i .

Llene el siguiente cuadro, analice los resultados y de una explicación de dicho comportamiento.

	(Q, Doc1)	(Q, Doc2)
DotProduct	35.31	27.45
cosine	0.52	0.62

P3. Esquematice el funcionamiento del índice invertido para dar soporte al modelo de ranked retrieval.



Función de recuperación usando la similitud de coseno:

```
COSINESCORE(q)
1  float Scores[N] = 0
2  float Length[N]
3  for each query term t
4  do calculate  $w_{t,q}$  and fetch postings list for t
5    for each pair(d,  $tf_{t,d}$ ) in postings list
6    do Scores[d] +=  $w_{t,d} \times w_{t,q}$ 
7  Read the array Length
8  for each d
9  do Scores[d] = Scores[d] / Length[d]
10 return Top K components of Scores[]
```

Grafique el índice invertido:

Indique los pasos de la construcción del índice: