Recuperación ACTO1 – SAR

(19/06/2017)

(IMPORTANTE: todos los cálculos se mostrarán truncados a dos decimales)

1) En una colección de test para una consulta tenemos 9 documentos relevantes. Entre los 12 documentos devueltos sólo 6 son relevantes ocupando las posiciones 2,3,6,8,9,12.

Se pide:

(a) Calcula la eficacia del sistema sin tener en cuenta el orden de los documentos en términos de Precisión, Recall, F-medida con β=1 (No se puntuarán las respuestas que consistan únicamente en el valor resultante) **(0,3 puntos)**

Precisión= 6/12 = 0,5 **Recall=** 6/9 = 0,66
F-medida= (2x0,5x0,66)/(0,5+0,66)=0,66/1.16=0,56

(b) Completa la Tablas de Precision y Recall Real y la Tabla de Precisión y Recall Interpolada. **(0,7 puntos)**

Tabla Precision&Recall Reales

	1	2	3	4	5	6	7	8	9	10	11	12
Relevante	no	yes	yes	no	no	yes	no	yes	yes	no	no	yes
Precisión	0	0,50	0,66	0,50	0,40	0,50	0,42	0,50	0,55	0,50	0,45	0,50
Recall	0	0.11	0.22	0.22	0.22	0.33	0.33	0.44	0.55	0.55	0.55	0.66

Tabla Precision&Recall Interpoladas

Precisión	1	0,66	0,66	0,55	0,55	0,55	0,5	0	0	0	0
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

2) Considérese la siguiente colección de 3 documentos:

Doc1: El desarrollo económico depende de las decisiones del gobierno
Doc2: El sector económico más importante impulsará el desarrollo económico
Doc3: La decisiones deben tomarse cuidadosamente por el gobierno

Se pide:

a. Completar la tabla considerando sólo los términos indicados en ella, considerando que los vectores de documentos tienen una frecuencia de términos log-pesado, idf y normalizado. **(0.7 puntos)**

Term			Doc1				Doc2				Doc3			
	df_t	idf_t	$tf_{t,d}$	peso (tf)	$w_{t,d}$=tfxidfd	L-Normaliz	$tf_{t,d}$	peso (tf)	$wt_{,d}$=tf $_x$idfd	L-Normaliz	$tf_{t,d}$	peso (tf)	$wt_{,d}$=tf $_x$idfd	L-Normaliz
desarrollo	2	0,17	1	1	0,17	0,50	1	1	0,17	0,31	0	0	0	0,00
económico	2	0,17	1	1	0,17	0,50	2	1,3	0,22	0,40	0	0	0	0,00
decisiones	2	0,17	1	1	0,17	0,50	0	0	0	0,00	1	1	0,17	0,70
gobierno	2	0,17	1	1	0,17	0,50	0	0	0	0,00	1	1	0,17	0,70
importante	1	0,47	0	0	0	0,00	1	1	0,47	0,86	0	0	0	0,00

b) Considerando la tabla anterior, qué par de documentos es más similar, Doc1 y Doc2,o Doc1 y Doc3? **(0.3 puntos)**

cos(Doc1, Doc2)= (0,5x0,31)+(0,5x0,4)+(0,5x0)+(0,5x0)+(0x0,86) = 0,35

cos(Doc1, Doc3)= (0,5x0)+(0,5x0)+(0,5x0,70)+(0,5x0,70)+(0x0) = 0,70

Por lo tanto el par (Doc1, Doc3) es más similar que el par de documentos (Doc1, Doc2).

3) Se pide escribir (en seudocódigo) el algoritmo que, a partir de una postings list correspondiente a la búsqueda del término A nos proporcionaría el resultado de la consulta (NOT A). La colección consta de N documentos y los DocID asignados son valores consecutivos entre 1 y N. El valor N es un parámetro del algoritmo. **(0,5 puntos)**

```
ALGORITMO NOT (p1, N)
respuesta ← {}
ID ← 1
mientras No_FINAL( p1)
hacer          mientras docID (p1) > ID
               hacer   Añadir (respuesta, ID)
                       ID ++
               ID ++
               p1 ← Avanzar_Siguiente(p1)
mientras ID < = N
hacer          Añadir (respuesta, ID )
               ID++
Return respuesta
```

4) Se pide completar la inserción en una tabla hash cerrada de tamaño B=12, con función hash H(x)= x MOD B, y con estrategia de redispersión 2ª función hash

hi(x) = (hi-1(x) + k(x)) MOD B siendo k (x) = (x MOD (B-2)) + 1

de los siguientes elementos: 25, 8, 37, 20, 49 **(0.5 puntos)**

0	
1	25
2	
3	
4	
5	
6	
7	
8	8
9	37
10	20
11	49