Recuperación ACTO1 – SAR

(18/06/2018 – 3 puntos)

Apellidos y Nombre:……

(IMPORTANTE: todos los cálculos se mostrarán truncados a dos decimales; se deben justificar las respuestas)

1) En una colección de test para una consulta tenemos 8 documentos relevantes. Entre los 10 documentos devueltos solamente 5 son relevantes ocupando las posiciones 1,3,5,8,9.

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= 5/10=0,5 **Recall=** 5/8= 0,62

F-medida= 2x0,5x0,62/(0,5+0,62)= 0,55

b) Completa las Tablas de Precision y Recall (expresando la operación de división realizada y el resultado en decimales, p.e. 2/3 = 0,66) e Interpoladas. **(0,7 puntos)**

Tabla Precision&Recall Reales

	1	2	3	4	5	6	7	8	9	10
Relevante	yes	no	yes	no	yes	no	no	yes	yes	no
Precisión	1	0,5	0,66	0,5	0,6	0,5	0,42	0,5	0,55	0,5
Recall	0,12	0,12	0,25	0,25	0,37	0,37	0,37	0,5	0,62	0,62

Tabla Precision&Recall Interpoladas

Precisión	1	1	0,66	0,6	0,55	0,55	0,55	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 4 documentos:

Doc1: El perro bebe agua de la mano de su dueño.
Doc2: El dueño de la cantina antes de dar de beber debería darle una mano de pintura a su local.
Doc3: Bebían y bebían como si no hubiesen bebido durante días o fueran los dueños del oasis.
Doc4: Agua que no has de beber, déjala correr.

Se pide:

a. Completar la tabla tomando como términos los lemas **beber**, **agua** y **dueño** representando las diversas palabras que corresponden a flexiones de estos lemas que aparecen en los documentos anteriores, considerando que se sigue un esquema log-pesado, idf y normalización coseno. **(0,6 puntos)**

Term			Doc1				Doc2				Doc3				Doc4			
	df_t	idf_t	$f_{t,q}$	tf_{tq}	$W_{t,q}$=tfxidf	L-Norm	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tf_xidf	L-Norm	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tf_xidf	L-Norm	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tf_xidf	L-Norm
agua	2	0,3	1	1	0,3	0,92	0	0	0	0,00	0	0	0	0,00	1	1	0,3	1,00
beber	4	0	1	1	0	0,00	1	1	0	0,00	3	1,47	0	0,00	1	1	0	0,00
dueño	3	0,12	1	1	0,12	0,37	1	1	0,12	1,00	1	1	0,12	1,00	0	0	0	0,00

b) Considerando la tabla anterior, qué otro documento es más similar a Doc1? **(0,4 puntos)**

Solución:

cos(Doc1, Doc2)= (0,92x0)+(0x0)+(0,37x1) = 0,37
cos(Doc1, Doc3)= (0,92x0)+(0x0)+(0,37x1) = 0,37
cos(Doc1, Doc4)= (0,92x1)+(0x0)+(0,37x0) = 0,92

Por lo tanto el Doc4 es el más similar al documento Doc1.

3) Esta pregunta consta de dos apartados: **(0,6 puntos)**

a) ¿Cómo sería el índice permuterm para la palabra "costa"?

Solución:
El índice permuterm para el término prisa se construiría con las diferentes rotaciones del término:

costa $
osta$c
sta$co
ta$cos
a$cost
$costa

b) ¿Cómo sería el mecanismo de búsqueda correspondiente a los wildcard queries "co*ta" y "*osta" suponiendo que disponemos de este tipo de índice?

Solución:

La búsqueda que se realiza para "co*ta" es "ta$co* y para "*osta" es "osta$*".

4) Se pide justificar la veracidad o falsedad de cada una de las siguientes afirmaciones: **(0,4 puntos)**

a) En un sistema de recuperación de información booleano usar stemming implica pérdida de precisión.
b) En un sistema de recuperación de información booleano usar stemming implica pérdida de recall.

Solución:

Hacer stemming comporta agrupar los términos que comparten el mismo stemming en un solo término tanto en el índice invertido como en la consulta. Tanto en la medida de precisión como en la de recall, el numerador es el número de documentos devueltos por el sistema que son correctos para la consulta.

a) Para el caso de la precisión el denominador es el número total de documentos devueltos, que en el caso de utilizar stemming será mayor que en el caso de no usarlo, por lo que el cociente resultante, la precisión será menor. Por tanto, la afirmación es cierta.
b) Para el caso del recall el denominador es el número de documentos de la referencia, que para una determinada consulta no varía se utilice stemming o no. Por tanto la afirmación es falsa ya que el recall será el mismo.