ACTO1 RECUPERACIÓN– SAR

(18/06/2021 – 2 puntos)

Apellidos y Nombre: ……

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

1. Sea una colección de documentos con 50 documentos, identificados con los números de 1 al 50. Sabemos que los documentos relevantes para una determinada consulta son [1,20,5,22,35,14,41,6].
Un sistema S de recuperación de información devuelve el siguiente resultado para la consulta:

S= [41,38, 7,1,5,23,44,6,21,14]

Se pide:

a) Calcular la eficacia (Precisión, Recall y la F-medida con β=1) para la consulta. **(0,2 puntos)**

Precisión	Recall	F-1

b) Completar las Tablas de Precision y Recall (expresando la operación de división realizada y el resultado redondeado en dos decimales, p.e. 2/3 = 0,67) e Interpoladas. **(0,6 puntos)**

Tabla Precision&Recall Reales

	1	2	3	4	5	6	7	8	9	10
Relevante										
Precisión										
Recall										

Tabla Precision&Recall Interpoladas

Precisión											
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

c) Calcular la Precisión Media

2. Considérese la siguiente colección de 4 documentos:

Doc1: Shared Computer Resources
Doc2: Computer Services
Doc3: Digital Shared Components
Doc4: Computer Resources Shared Components

Asumiendo que cada palabra es un término se pide:

a. Escribir la matriz de incidencia binaria. **(0.1 punto)**

	Doc1	Doc2	Doc3	Doc4
Shared				
Computer				
Resources				
Services				
Digital				
Components				

b. Qué documentos serán recuperados, en el modelo booleano, con la consulta "Computer AND NOT Components"? (justifíquese la respuesta) **(0.2 puntos)**

c. Calcular la similitud entre la consulta "Computer Components" y Doc4, usando la similitud coseno y utilizando el esquema de pesado lnc.ltc. Puedes ayudarte de la tabla siguiente.

(0.5 puntos)

Term			Consulta				Doc4			
	df_t	idf_t	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tfxidfd	L-Norm	$f_{t,d}$	$tf_{t,d}$	wt,d=tf_xidfd	L-Norm
Shared										
Computer										
Resources										
Services										
Digital										
Components										

3. Se pide responder las siguientes preguntas: **(0.4 puntos)**

a) Describe en qué consiste, para qué sirve, y cómo se construye un índice Permuterm.
b) ¿Cuál sería el mecanismo de búsqueda correspondiente a los wildcard queries "h*ón" y "*urón" suponiendo que disponemos de este tipo de índice?

Soluciones:

1. Sea una colección de documentos con 50 documentos, identificados con los números de 1 al 50. Sabemos que los documentos relevantes para una determinada consulta son [1,20,5,22,35,14,41,6].
Un sistema S de recuperación de información devuelve el siguiente resultado para la consulta:

S= [**41**,38, 7,**1,5**,23,44,**6**,21,**14**]

Se pide:

a) Calcular la eficacia (Precisión, Recall y la F-medida con β=1) para la consulta. **(0,2 puntos)**

Precisión	Recall	F-1
5/10=0,5	5/8=0,63	O,56

b) Completar las Tablas de Precision y Recall (expresando la operación de división realizada y el resultado redondeado en dos decimales, p.e. 2/3 = 0,67) e Interpoladas. **(0,6 puntos)**

Tabla Precision&Recall Reales

	1	2	3	4	5	6	7	8	9	10
Relevante	Y	N	N	Y	Y	N	N	Y	N	Y
Precisión	1	0,5	0,33	0,5	0,6	0,5	0,43	0,5	0,44	O,5
Recall	0,13	O,13	0,13	0,25	0,38	0,38	0,38	0,5	0,5	0,63

Tabla Precision&Recall Interpoladas

Precisión	1	1	0,6	0,6	0,5	0,5	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

c) Calcula la Precisión Media (1+0,5+0,6+0,5+0,5+0+0+0)/8=0,39

2.

a. Escribir la matriz de incidencia binaria **(0.1 punto)**

	Doc1	Doc2	Doc3	Doc4
Shared	1	0	1	1
Computer	1	1	0	1
Resources	1	0	0	1
Services	0	1	0	0
Digital	0	0	1	0
Components	0	0	1	1

b. Qué documentos serán recuperados, en el modelo booleano, con la consulta: "Computer AND NOT Components"? (justifíquese la respuesta) **(0.2 puntos)**

Computer= 1101
Components= 0011
Computer AND NOT Components = 1101 AND NOT 0011= 1101 AND 1100 = 1100
Por tanto, la respuesta es: Doc1 y Doc2

c. Calcular la similitud entre la consulta "Computer Components" y Doc4, usando la similitud coseno y utilizando el esquema de pesado lnc.ltc. Puedes ayudarte de la tabla siguiente. **(0.5 puntos)**

Term			Consulta				Doc4			
	df_t	idf_t	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}=tf_x idfd$	L-Normaliz	$f_{t,d}$	$tf_{t,d}$	$wt,d=tf_x idfd$	L-Normaliz
Shared	3	0,12	0	0	0	0,00	1	1	1	0,50
Computer	3	0,12	1	1	0,12	0,37	1	1	1	0,50
Resources	2	0,3	0	0	0	0,00	1	1	1	0,50
Services	1	0,6	0	0	0	0,00	0	0	0	0,00
Digital	1	0,6	0	0	0	0,00	0	0	0	0,00
Components	2	0,3	1	1	0,3	0,93	1	1	1	0,50

$$tf_{t,d} = \begin{cases} 1 + \log_{10} f_{t,d}, & \text{si } f_{t,d} > 0 \\ 0, & \text{otro caso} \end{cases}$$

$$\text{idf}_t = \log_{10}(N/\text{df}_t)$$

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|k|} q_i d_i}{\sqrt{\sum_{i=1}^{|k|} q_i^2}\sqrt{\sum_{i=1}^{|k|} d_i^2}}$$

Por tanto la solución es cos(q,d4)=(0,37x0,5)+(0,93x0,5)=0,66

3. **(0.4 puntos)**

a) Comenta brevemente en qué consiste y cómo se construye un índice Permuterm

Un índice Permuterm es un segundo índice que se construye para poder hacer búsquedas con tolerancia. A cada término se añade un símbolo $ al final, y se inserta una entrada al índice para cada rotación del término que enlaza con el término original. Para la wildcardquery m*n se rotaría apareciendo el * al final de la cadena buscando la cadena n$m* en el Índice Permuterm

b) ¿Cómo sería el mecanismo de búsqueda correspondiente a los wildcard queries “h*ón” y “*urón” suponiendo que disponemos de este tipo de índice?

Aplicando el mecanismo indicado en el apartado anterior la búsqueda en el índice Permuterm se realizaría con: ón$h* y urón$*