ACTO1 – SAR

(27/03/2017)

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

1) Sea una colección de documentos con 30 documentos, identificados con los números de 1 al 30. Sabemos que los documentos relevantes para una determinada consulta son los numerados de1 1 al 10.

Dos sistemas de recuperación de información devuelven el siguiente resultado para la consulta:

S1= [1,11,2,12,13,3,14,4,15,16,27,5]
S2= [11,12,1,13,2,14,24,16,3,17,4,18]

Para cada uno de los sistemas se pide:

a) Calcular la eficacia (Precisión, Recall y la F-medida con β=1) para la consulta. **(0,3 puntos)**

P(S1)=5/12=0.41 R(S1)=5/10= 0.5 F1= 2(0.41x0.5)/(0.41+0.5)=0.45
P(S2)=4/12=0.33 R(S2)=4/10=0.4 F1= 2(0.33x0.4)/(0.33+0.4)=0.35

b) Completa las Tablas de Precision y Recall Reales (expresando la operación de división realizada y el resultado en decimales, p.e. 2/3 = 0,66) e Interpoladas. **(0,9 puntos)**

Tabla Precision&Recall Reales

S1	1	2	3	4	5	6	7	8	9	10	11	12
Relevante	yes	no	yes	no	no	yes	no	yes	no	no	no	yes
Precisión	1/1	1/2	2/3	2/4	2/5	3/6	3/7	4/8	4/9	4/10	4/11	5/12
Recall	0.1	0.1	0.2	0.2	0.2	0.3	0.3	0.4	0.4	0.4	0.4	0.5

S2	1	2	3	4	5	6	7	8	9	10	11	12
Relevante	no	no	yes	no	yes	no	no	no	yes	no	yes	no
Precisión	0	0	1/3	1/4	2/5	2/6	2/7	2/8	3/9	3/10	4/11	4/12
Recall	0	0	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.4	0.4

Tabla Precision&Recall Interpoladas

Precisión S1	1	1	2/3	3/6	4/8	5/12	0	0	0	0	0
Precisión S2	1	2/5	2/5	4/11	4/11	0	0	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) Teniendo en cuenta los resultados de los anteriores apartados, ¿cuál de los dos sistemas es mejor? (justifíquese la respuesta) **(0,2 puntos)**

El S1, presenta unos resultados mejores tanto en Precisión como en Recall, y por tanto en F1. Observamos también un mejor comportamiento en la gráfica interpolada, ya que S1 devuelve más resultados relevantes en las primeras posiciones.

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 y que representamos los términos por orden de aparición en la colección, se pide:

a. Escribir la matriz de incidencia binaria. **(0.2 puntos)**

	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. Puedes ayudarte de la tabla siguiente. **(0.6 puntos)**

Term	Consulta						Doc4				Producto
	$tf_{t,q}$	peso($tf_{t,q}$)	df_t	idf_t	$W_{t,q}$=tfxidf	L-Normaliz	$tf_{t,d}$	peso (tf)	$w_{t,d}$=tf	L-Normaliz	
Shared	0	0	3	0,12	0	0,00	1	1	1	0,50	0,00
Computer	1	1	3	0,12	0,12	0,37	1	1	1	0,50	0,18
Resources	0	0	2	0,3	0	0,00	1	1	1	0,50	0,00
Services	0	0	1	0,6	0	0,00	0	0	0	0,00	0,00
Digital	0	0	1	0,6	0	0,00	0	0	0	0,00	0,00
Components	1	1	2	0,3	0,3	0,93	1	1	1	0,50	0,46

Por tanto la solución es cos(q, d)= 0+0,18+0+0+0+0,46 = 0,64

Utilizamos el esquema de pesado lnc.ltc:

- para los documentos log-pesado, no idf y coseno normalizado;
- para la consulta log-pesado, idf y coseno normalizado.

3)

a) Se quiere hacer un búsqueda con tolerancia de las consultas ca*na y *mente. Qué término ha de buscarse suponiendo que tenemos un diccionario con entradas “permuterm”? **(0.3 puntos)**

na$ca*
mente$*
Por tanto, en el diccionario del índice permuterm se buscaran los términos que empiecen por “na$ca” y “mente$” respectivamente.

b) Sea el siguiente diccionario de bigramas. Indica qué términos devolvería para la consulta ca*na. Indica qué bigramas están implicados en la consulta. **(0.3 puntos)**

$a ➡	acaba	anaconda								
$c ➡	camina	cana	cansado	canto	casa	cena	cesta	comida	concatena	
a$ ➡	anaconda	acaba	brinca	camina	cana	casa	cena	cesta	comida	concatena
an ➡	anaconda	cana	cansado	canto						
ca ➡	acaba	brinca	camina	cana	canto	casa	concatena			
na ➡	anaconda	camina	cana	cena	concatena					

camina, cana, concatena
Resulta de hacer $c AND ca AND na AND a$
El resultado “concatena” no es una buena respuesta e ilustra una desventaja del método de los bigramas.