ACTO1 – SAR

(25/03/2019)

Apellidos y Nombre:……

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

1) Sea una colección de documentos con 40 documentos, identificados con los números de 1 al 40. Sabemos que los documentos relevantes para una determinada consulta son [2,5,7,14,15,16,19,23,34,39].
Dos sistemas de recuperación de información devuelven el siguiente resultado para la consulta:

S1= [5, 11, 2, 19, 14, 3, 35, 34, 33, 16, 1, 8]
S2= [34, 1, 7, 19, 12, 20, 24, 16, 3, 17, 33, 18]

Para cada uno de los sistemas se pide: **(1 punto)**

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

Consulta	Precisión	Recall	F-1
S1	6/12=0.5	6/10=0.6	0.54
S2	4/12=0.33	4/10=0.4	0.36

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

Tabla Precision&Recall Reales

S1	1	2	3	4	5	6	7	8	9	10	11	12
Relevante	Y	N	Y	Y	Y	N	N	Y	N	Y	N	N
Precisión	1	0.5	0.66	0.75	0.8	0.66	0.57	0,62	0.55	0.6	0.54	0.5
Recall	0.1	0.1	0.2	0.3	0.4	0.4	0.4	0.5	0.5	0.6	0.6	0.6

S2	1	2	3	4	5	6	7	8	9	10	11	12
Relevante	Y	N	Y	Y	N	N	N	Y	N	N	N	N
Precisión	1	0.5	0.66	0.75	0.6	0.5	0.42	0.5	0.44	0.4	0.36	0.33
Recall	0.1	0.1	0.2	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.4

Tabla Precision&Recall Interpoladas

Precisión S1	1	1	0.8	0.8	0.8	0.62	0.6	0	0	0	0
Precisión S2	1	1	0.75	0.75	0.5	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) Calcula la precisión media para cada uno de los sistemas.

P_MEDIA_S1=(1+0.66+0.75+0.8+0.62+0.6+0+0+0+0)/10= 4.43/10= 0.44
P_MEDIA_S2=(1+0.66+0.75+0.5+0+0+0+0+0+0)/10= 2.91/10 = 0.29

2) Considérese la siguiente colección de 3 documentos:

Doc1: **pronto**, muy **pronto**, se va a **iniciar** el **mundial**
Doc2: **pronto** también se va a **iniciar** el **tour** de **Francia**
Doc3: se va a **iniciar** un **torneo** de **tenis** en **Francia**

Los términos a considerar se han indicado en negrita.
Se pide: **(1 punto)**

a) Completar la tabla considerando que se sigue un esquema log-pesado, idf y normalización coseno.
b) A partir de la tabla, calcular las similitudes entre los 3 pares de documentos ¿qué par de documentos es más similar?

Solución:

Term			Doc1				Doc2				Doc3			
	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
pronto	2	0,17	2	1,3	0,22	0,42	1	1	0,17	0,32	0	0	0	0,00
iniciar	3	0	1	1	0	0,00	1	1	0	0,00	1	1	0	0,00
mundial	1	0,47	1	1	0,47	0,90	0	0	0	0,00	0	0	0	0,00
tour	1	0,47	0	0	0	0,00	1	1	0,47	0,89	0	0	0	0,00
Francia	2	0,17	0	0	0	0,00	1	1	0,17	0,32	1	1	0,17	0,24
torneo	1	0,47	0	0	0	0,00	0	0	0	0,00	1	1	0,47	0,68
tenis	1	0,47	0	0	0	0,00	0	0	0	0,00	1	1	0,47	0,68

cos(Doc1,Doc2)= (0,42x0,32)+(0x0)+(0,9x0)+(0x0,89)+(0x0,32)+(0x0)+(0x0)=0,13
cos(Doc1,Doc3)= (0,42x0)+(0x0)+(0,9x0)+(0x0)+(0x0,24)+(0x0,68)+(0x0,68)=0
cos(Doc2,Doc3)= (0,32x0)+(0x0)+(0x0)+(0,89x0)+(0,32x0,24)+(0x0,68)+(0x0,68)=0,07

Por tanto los documentos más similares son Doc1 y Doc2

3) Se pide escribir (en seudocódigo) el algoritmo que, a partir de las postings lists correspondientes a la búsqueda de los términos A y B, nos proporcionaría el resultado del Query: A OR B.: **(0,4 puntos)**

Suponemos las postings lists ordenadas por docID.
Sean p1 y p2 los punteros al principio de dichas postings lists

```
ALGORITMO OR (p1, p2)
respuesta ← {}
mientras No_FINAL( p1) AND No_FINAL( p2)
hacer    si docID (p1) = docID (p2)
         entonces Añadir (respuesta, docID (p1))
                  p1 ← Avanzar_Siguiente(p1)
                  p2 ← Avanzar_Siguiente(p2)
         sino     si docID (p1) < docID (p2)
                  entonces  Añadir (respuesta, docID (p1))
                            p1 ← Avanzar_Siguiente(p1)
                  sino      Añadir (respuesta, docID (p2))
                            p2 ← Avanzar_Siguiente(p2)
mientras No_FINAL( p1)
hacer    Añadir (respuesta, docID (p1))
         p1 ← Avanzar_Siguiente(p1)
mientras No_FINAL( p2)
hacer    Añadir (respuesta, docID (p2))
         p2 ← Avanzar_Siguiente(p2)
```

4) Se pide responder las siguientes preguntas: **(0,6 puntos)**

a) Comenta brevemente en qué consiste y cómo se construye un índice de n-gramas.
b) Explica cómo sería la búsqueda de documentos correspondientes a la wildcard query "ca*sa".
c) Si tenemos el siguiente diccionario de bigramas indica que términos devolvería para la consulta ca*sa. Comenta también si todas las palabras devueltas son correctas para la consulta realizada.

$a ➡	acabo	antena	antigua	asar					
$c ➡	camino	comino	camisa	canto	cansa	cena	comida	carcasa	casaca
a$ ➡	cansa	antena	antigua	camisa	carcasa	poca	casaca	comida	cena
an ➡	antigua	cansa	pantano	canto	antena	gusano			
ca ➡	acabo	camisa	cansa	casaca	canto	carcasa	rocas	poca	camino
sa ➡	pasar	carcasa	cansa	pesar	camisa	cosas	casaca	asar	gusano

a. Un índice de n-gramas es un segundo índice que se construye para poder hacer búsquedas con tolerancia. Se calculan los n-gramas de caracteres a partir de los términos que aparecen en los documentos a los que se ha añadido previamente el símbolo "$" al inicio y fin. En el índice de n-gramas, cada n-grama apunta a la lista de términos que contienen ese n-grama.

b. La búsqueda de documentos correspondientes a la wildcard query "ca*sa"se realizaría a partir de la expresión lógica: $c AND ca AND sa AND a$ que utilizando un algoritmo de INTERSECCIÓN de las listas de términos correspondientes a cada bigrama nos devolvería la lista de términos resultante.

c. Términos que devolvería: "camisa", "carcasa", "cansa", "casaca". El término "casaca" es incorrecto para la consulta realizada ya que no acaba en "sa", pero sería devuelto por el índice.