ACTO1 – SAR

(29/03/2021 – 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 100 documentos, identificados con los números de 1 al 100. Sabemos que los documentos relevantes para una determinada consulta son [3,5,18,22,35,40,41,63,80,89].

Un sistema S de recuperación de información devuelve el siguiente resultado para la consulta:

S= [22, 81, 3, 26, 40, 7, 80, 5, 25, 19, 89, 76]

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	11	12
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

2. Sean una colección de documentos compuesta únicamente por los documentos Doc1 y Doc2 y sea la siguiente consulta:

Doc1: El **desarrollo económico** depende de las **decisiones** del **gobierno**
Doc2: El **sector económico** más importante impulsará las **decisiones económicas**

consulta: Las **decisiones económicas** deben tomarse por el **gobierno**

Los términos a considerar se han indicado en negrita, se ha llevado a cabo lematización tanto en consulta como en documentos, de forma que por ejemplo "económico" y "económicas" se consideran el mismo término.

Se pide:

a) Completar la tabla para un esquema de pesado ltc.ltc (log-pesado, idf y coseno normalizado). **(0,3 puntos)**

Term			Consulta				Doc1				Doc2			
	df_t	idf_t	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tfxidfd	L-Normaliz	$f_{t,d}$	$tf_{t,d}$	wt,d=tfxidfd	L-Normaliz	$f_{t,d}$	$tf_{t,d}$	wt,d=tfxidfd	L-Normaliz
desarrollo														
económico														
decisión														
gobierno														
sector														

b) Indicar qué documento es más relevante para la consulta en base a la similitud coseno con esquema de pesado ltc.ltc. **(0,2 puntos)**

3. En un índice invertido construido para una colección de **N** documentos, se realiza la inserción de un nuevo documento **d**. Se pide:

a) ¿Qué acciones deberían realizarse sobre el diccionario y las listas de postings del índice para todo término **t** del documento **d**? **(0,2 puntos)**

b) ¿Qué valores deberían crearse y/o actualizarse para poder aplicar el modelo vectorial con un esquema de pesado tf-idf para todo término **t** del documento **d**? **(0,2 puntos)**

4. Se pide calcular la distancia de Levenshtein entre las palabras **oxus y ohxoos,** considerando que el coste de la operación Borrado es 1, Inserción es 1, y Sustitución es 1. Utiliza la cuadrícula para representar los costes acumulados. La cuadrícula tiene un tamaño fijo, que no tiene por qué ajustarse exactamente al espacio que se requiere. **(0.3 puntos)**

Soluciones:

1) Sea una colección de documentos con 100 documentos, identificados con los números de 1 al 100. Sabemos que los documentos relevantes para una determinada consulta son [3,5,18,22,35,40,41,63,80,89].

Un sistema S de recuperación de información devuelve el siguiente resultado para la consulta:

S1= [22, 81, 3, 26, 40, 7, 80, 5, 25, 19, 89, 76]

Se pide:

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

$$F_1 = \frac{2PR}{P+R}$$

Precisión	Recall	F-1
6/12=0,5	6/10=0,6	0,55

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 = 0,67) e Interpoladas.

Tablas Precision&Recall Reales

	1	2	3	4	5	6	7	8	9	10	11	12
Relevante	Y	N	Y	N	Y	N	Y	Y	N	N	Y	N
Precisión	1/1=1	½=0,5	2/3=0,67	2/4=0,5	3/5=0,6	3/6=0,5	4/7=0,57	5/8=0,63	5/9=0,56	5/10=0,5	6/11=0,55	6/12=0,5
Recall	1/10=0,1	1/10=0,1	2/10=0,2	2/10=0,2	3/10=0,3	3/10=0,3	4/10=0,4	5/10=0,5	5/10=0,5	5/10=0,5	6/10=0,6	6/10=0,6

Tablas Precision&Recall Interpoladas

Precisión	1	1	0,67	0,63	0,63	0,63	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

Para cada valor de recall estándar **i** desde 0.0 a 1.0 con incrementos de 0.1, se toma la precisión máxima obtenida en cualquier valor de recall real mayor o igual a **i**.

2.

2.a)

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

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

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

Term			Consulta				Doc1				Doc2			
	df_t	idf_t	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tfxidfd	L-Normaliz	$f_{t,d}$	$tf_{t,d}$	$wt_{,d}$=tf_xidfd	L-Normaliz	$f_{t,d}$	$tf_{t,d}$	wt,d=tfxidfd	L-Normaliz
desarrollo	1	0,3	0	0	0	0,00	1	1	0,3	0,71	0	0	0	0
económico	2	0	1	1	0	0,00	1	1	0	0,00	2	1,3	0	0
decisión	2	0	1	1	0	0,00	1	1	0	0,00	1	1	0	0
gobierno	1	0,3	1	1	0,3	1,00	1	1	0,3	0,71	0	0	0	0
sector	1	0,3	0	0	0	0,00	0	0	0	0,00	1	1	0,3	1

2.b)

cos(q,doc1)= (1x0,71)=0,71

cos(q,doc2)= 0

Por lo que para la consulta es más relevante Doc1

3.

3.a)

Para cada término **t** del nuevo documento **d**:

- si el término ya existe en el diccionario, hay que localizar su entrada en él y añadir a su lista de postings una nueva entrada para el documento **d**.
- si el término **t** no existe en el diccionario, hay que añadirlo, crear la lista de postings para el término **t** y añadirle una nueva entrada para el documento **d**.

3.b)

El número total de documentos de la colección pasará a ser **N+1,** por lo que habrá que tenerlo en cuenta para el cálculo todos los **idf**.

Para cada término **t** del nuevo documento **d** habrá que actualizar el valor $\mathbf{df_t}$.

Para cada nuevo posting de los términos **t** del nuevo documento **d** hay que añadir la frecuencia del término en el documento $\mathbf{f_{td}}$.

4.

s	4	3	3	3	3	3	3		
u	3	2	2	2	2	3	4		
x	2	1	1	1	2	3	4		
o	1	0	1	2	3	4	5		
#	0	1	2	3	4	5	6		
	#	o	h	x	o	o	s		

La distancia de Levenshtein es 3, valor de la esquina superior derecha de la tabla.