ACTO1 – SAR

(29/03/2021 – 2 puntos)

Apellidos y Nombre: David Arnal García
(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 = [18, 22, 93, 40, 4, 7, 6, 63, 62, 19, 2, 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
Rr / recu = 4 / 12 = **0,33**	Rr / R = 4 / 10 = **0,4**	2PR / P + R = **0,36**

N = 100
R = 10
Rr = 4
recu = 12

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	Yes	Yes	No	Yes	No	No	No	Yes	No	No	No	No
Precisión	1 / 1 = 1	2 / 2 = 1	2 / 3 = 0,67	3 / 4 = 0,75	3 / 5 = 0,6	3 / 6 = 0,5	3 / 7 = 0,43	4 / 8 = 0,5	4 / 9 = 0,44	4 / 10 = 0,4	4 / 11 = 0,36	4 / 12 = 0,33
Recall	0,1	0,2	0,2	0,3	0,3	0,3	0,3	0,4	0,4	0,4	0,4	0,4

1 / R = 1 / 10 = 0,1

Tabla Precision&Recall Interpoladas

Precisión	1	1	1	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

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

Doc1: donde hay **vino beben vino**, donde no hay **vino agua fresca**
Doc2: le **gustaba beber vino** como si de **agua** se tratara

consulta: me **gusta** el **agua fresca**

Los términos a considerar se han indicado en negrita, se ha realizado una lematización, por lo que por ejemplo 'beben' y 'bebía' se representan con el término 'beber'.

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	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tf.idf	L-Norm	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tf.idf	L-Norm	$f_{t,d}$	$tf_{t,d}$	$w_{t,d}$=tf.idf	L-Norm
vino	2	0	0	0	0	0	3	1,48	0	0	1	1	0	0
beber	2	0	0	0	0	0	1	1	0	0	1	1	0	0
agua	2	0	1	1	0	0	1	1	0	0	1	1	0	0
fresca	1	0,3	1	1	0,3	0,71	1	1	0,3	1	0	0	0	0
gustar	1	0,3	1	1	0,3	0,71	0	0	0	0	1	1	0,3	1

$N = 2$

$idf_t = \log_{10}(N / df_t)$

$tf_{t,d} = 1 + \log_{10}(f_{t,d})$

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)**

cos (consulta, Doc1) = 0 * 0 + 0 * 0 + 0 * 0 + 0,71 * 1 + 0,71 * 0 = **0,71.**

cos (consulta, Doc2) = 0 * 0 + 0 * 0 + 0 * 0 + 0,71 * 0 + 0,71 * 1 = **0,71.**

Como se comprueba matemáticamente, **ambos documentos serían igual de relevantes para la consulta**.

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)**

El diccionario debería añadir los nuevos términos del documento *d* y, respecto a las *posting lists*, se deberían actualizar para todos los términos añadiendo en sus *posting list* la ocurrencia con el nuevo documento, si fuera el caso.

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)**

Se deberían añadir los nuevos términos del documento *d* y añadir su t_f y calcular sus idf_t. Además, habría que recalcular todos los idf_t de todos los términos, ya que depende de todos los documentos.

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)**

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

D (oxus, ohxoos) = 3