Examen **Recuperación de Información** Julio 2016

Apellidos: _____Nombre: _____

Examen sin libros, apuntes, ni dispositivos electrónicos.
Tiempo: 2h

1. (1.5 puntos)
Pseudocódigo, estructuras de datos y funciones, explicación. Debe contestar a cada apartado en el espacio dedicado.

a) (0.6) Pseudocódigo del hilo de crawler estudiado en clase.

Slide 9 del tema de crawling. El pseudocódigo debe ser correcto y cumplir con la funcionalidad de un hilo de crawling. Obviamente no tiene que ser exacto al de las slides

b) (0.2) Explicación de estructuras de datos y funciones principales usadas en a)

Se trata de en a) contestar sólo con el pseudocódigo y aquí explicar las estructuras de datos funciones principales.

c) (0.2) Explicación en lenguaje natural

Una explicación breve en lenguaje natural facilita la compresión del ejercicio y la tarea del profesor que evalúa.

d) (0.5) ¿Puede adaptarse ese código para que el crawler visite las páginas con una estrategia primero en anchura y de forma que la profundidad en cada sitio web esté limitada a un valor de 3? Marque la respuesta correcta y razónelo.
Si es posible diga como, si no es posible diga por qué.

-Es posible

-No es posible

Si es posible. Si la frontera se maneja como una cola FIFO, resultará en que se visitan las páginas con un orden de primero en anchura. Cuando se parsea una página web para extraer los enlaces, habría que analizar estos enlaces y no meterlos en la frontera si corresponden a profundidades mayores de 3.

2. (1.5 puntos) Pseudocódigo, estructuras de datos y funciones, explicación. Debe contestar a cada apartado en el espacio dedicado.

a) (0.6) Pseudocódigo del algoritmo Term-At-A-Time de procesamiento de consultas estudiado en clase.

Slide 41 de capítulo de Ranking with Indexes. El pseudocódigo debe ser correcto y cumplir con la funcionalidad de algoritmo TAAT. Obviamente no tiene que ser exacto al de las slides

b) (0.2) Explicación de estructuras de datos y funciones principales usadas en a)

Se trata de en a) contestar sólo con el pseudocódigo y aquí explicar las estructuras de datos funciones principales.

c) (0.2) Explicación en lenguaje natural

Una explicación breve en lenguaje natural facilita la compresión del ejercicio y la tarea del profesor que evalúa.

d) (0.5 puntos) ¿Puede adaptarse ese código para procesar consultas en el modelo de RI basado en modelos de lenguaje Query Likelihood con suavización de Dirichlet? Marque la respuesta correcta y razónelo. Si es posible diga como, si no es posible diga por qué.

-Es posible

-No es posible

Si. Ya que VSM y Language Models (al computar log(P(q|d)) tiene una forma analítica similar (un sumatorio de expresiones donde aparecen los pesos que se obtienen con información almacenada o bien en el lexicon (df, cf), en las listas invertidas (tf(t,d)) o en otro fichero (|d|, |C|). Por tanto gi(Q) x f(li) en VSM son lo pesos del query term en la query y en el documeto. En el caso de Language Models hay tendriamos el log de la probabilidad suavizada del query term en el documento. Para Language Models necesitamos |d|, pero necesitamos ademas |C| y cf(t) en vez de df(t). Por tanto en indexación tienen que obtenerse esos valores pero el esqueleto de TAAT es el mismo y los dos modelos pueden trabajar sobre el índice invertido.

3. (1 punto) Una query que tiene como total de documentos relevantes los documentos con docID: 3, 6, 7, 12, 13, y un sistema de RI recupera el siguiente ranking de documentos: 2, 6, 7, 8, 3, 12, 11, 14, 5, 13, donde de nuevo los números son los docID de los documentos.

a) (0.4) Rellene la tabla mostrando el recall y la precisión calculada para cada documento recuperado, es decir en cada posición del ranking.

Ranking	Recall	Precisión
d2	0	0
d6 R	1/5=0.2	0.5
d7 R	2/5= 0.4	2/3= 0.67
d8	0.4	0.5
d3 R	3/5= 0.6	3/5=0.6
d12 R	4/5 = 0.8	4/6= 0.67
d11	0.8	4/7=0.57
d14	0.8	4/8 = 0.5
d5	0.8	4/9 = 0.44
d13 R	5/5 = 1	0.5

b) (0.4) Rellene la tabla con la precisión interpolada a los niveles estándar de recall.

Niveles estándar de Recall	Precisión Interpolada
0.0	0.67
0.1	0.67
0.2	0.67
0.3	0.67
0.4	0.67
0.5	0.67
0.6	0.67
0.7	0.67
0.8	0.67
0.9	0.5
1.0	0.5

c) (0.2) Compute el MAP

(0.5+0.67+0.6+0.67+0.5)/5

4. (0.5 puntos) Considere una colección de documentos que contiene estos documentos

d1: scary green crocodile
d2: scary green big
d3: small crocodile

y la query: big green crocodile

Considere el esquema de pesado rawtf x idflog, donde los log son base 2, tanto en la query como en los documentos.

Compute la similaridad de coseno para esa query de los tres vectores y el ranking producido

Terms & df	idflog2	raw tf q	tf x idflog2 q	raw tf d1	tf x idflog2 d1
scary 2	0.58	0	0	1	0.58
green 2	0.58	1	0.58	1	0.58
crocodile 2	0.58	1	0.58	1	0.58
big 1	1.58	1	1.58	0	0
small 1	1.58	0	0	0	0

$|q| = \sqrt{0.58^2+0.58^2+1.58^2}=1.78$ $|d1|= \sqrt{0.58^2+0.58^2+0.58^2} = 1.00$
$\cos(q, d1) = (0.58 \times 0.58 + 0.58 \times 0.58)/ (1.78 \times 1.00) = 0.38$

Terms & df	idflog2	raw tf q	tf x idflog2 q	raw tf d2	tf x idflog2 d2
scary 2	0.58	0	0	1	0.58
green 2	0.58	1	0.58	1	0.58
crocodile 2	0.58	1	0.58	0	0
big 1	1.58	1	1.58	1	1.58
small 1	1.58	0	0	0	0

$|d2| = \sqrt{0.58^2+0.58^2+1.58^2}= 1.78$ $\cos(q, d2) = (0.58 \times 0.58 + 1.58 \times 1.58) / (1.78 \times 1.78) = 0.89$

Terms & df	idflog2	raw tf q	tf x idflog2 q	raw tf d3	tf x idflog2 d3
scary 2	0.58	0	0	0	0
green 2	0.58	1	0.58	0	0
crocodile 2	0.58	1	0.58	1	0.58
big 1	1.58	1	1.58	0	0
small 1	1.58	0	0	1	1.58

$|d3| = \sqrt{0.58^2+1.58^2}= 1.68$ $\cos(q, d3) = (0.58 \times 0.58) / (1.78 \times 1.68) = 0.11$

5. (0.5 puntos) Explique la fórmula de Rocchio para Relevance Feedback. Dada una query expresada por el vector q = <2, 0, 3, 1, 0> y los documentos devueltos por el sistema y marcados como relevantes d1, d2, d3 y como irrelevantes d4, d5, compute la query modificada supuesto el mismo peso para la query original, los documentos relevantes y los irrelevantes.

d1 = <3, 1, 2, 1, 0>
d2 = <4, 1, 3, 2, 2>
d3 = <1, 0, 5, 0, 3>
d4 = <1, 3, 0, 1, 2>
d5 = <0, 4, 0, 2, 2>

Relevance Feedback para VSM con Rocchio lo explicamos en el tema de Retrieval Models, slide 14 de ese tema. Para los datos de este problema tenemos que tomar $\alpha=\beta=\gamma=1$

Por tanto

query expandida=<2, 0, 3, 1, 0> + (⅓) (<3, 1, 2, 1, 0> + <4, 1, 3, 2, 2> + <1, 0, 5, 0, 3>) -
(½) (<1, 3, 0, 1, 2> + <0, 4, 0, 2, 2>) =
<2, 0, 3, 1, 0> +
<2.67, 0.67 , 3.33, 1, 1.67> -
<0.5, 3.5, 0, 1.5, 2> =
= <4.17, -2.83, 6.33, 0.5, -0.33>

VSM (Vector Space Model) no puede manejar vectors con componentes negativos, por tanto en la práctica la query expandida sería <4.17, 0, 6.33, 0.5, 0 >