Examen **Recuperación de Información** Junio 2021

Apellidos: _____Nombre: _____

Examen sin libros, apuntes, ni dispositivos electrónicos (sólo calculadora).
Tiempo: 2h. 15m.

EN LAS PREGUNTAS 1, 2, 3, 4 TIENEN QUE ESTAR CORRECTOS LOS CÁLCULOS, RESULTADOS Y EXPLICACIONES.

1. (1 punto) Considere un grafo web con 4 nodos (1, 2, 3, 4) y los siguientes enlaces. Del nodo 1 sale un enlace a 2 y otro a 4. Del nodo 3 sale un enlace a 2 y otro a 4. Del nodo 2 sale un enlace a 2. Del nodo 4 sale un enlace a 4. Compute la matriz de transición de probabilidad con un teleporting del 40% (MTP40%) y el Page Rank.

(MTP40%) :

0.1 0.4 0.1 0.4
0.1 0.7 0.1 0.1
0.1 0.4 0.1 0.4
0.1 0.1 0.1 0.7

Page Rank: 0.1 0.4 0.1 0.4

2. (1 punto) Considere relevancia binaria y una query con un total de 5 relevantes y un ranking (de izquierda a derecha las posiciones del ranking son de la 1 a la 10 y la relevancia de un documento en una posición se indica por R, la no relevancia por N) :

R R N R R R N N N N

a) (0.3) Compute MAP con corte 5
Respuesta: _____

(1 + 1 + 0.75 + 0.8) / 5 = 0.71

b) (0.3) Compute F1 (F con beta=1) con P y R computados con corte 5
Respuesta _____

F1= 2PR / (P+R) = 2 * 0.8 * 0.8 /(0.8 + 0.8) = 0.8

c) (0.4) Compute NDCG@5 con la siguiente formulación de DCG que sigue. (log2 N = (log10 N/ log10 2))

$$DCG@p = rel_1 + \sum_{i=2}^{p} \left(rel_i / \log_2 i \right)$$

Respuesta: _____
Dado el ranking
R R N R R R N N N N
La secuencia rel1, reli/log 2 i, con i=2,...5 es:
1 1/1 0/1.58 ½ 1/2.32
1 1 0 0.5 0.43
Por tanto DCG@p con p=1,...5 es
1 2 2 2.5 2.93
Para el ranking ideal:
R R R R R N N N N N
La secuencia rel1, reli/log 2 i con i=2,...10 es:
1 1/1 1/1.58 ½ 1/2.32
1 1 0.63 0.5 0.43
Por tanto DCG@p con p=1,...5 para el ranking ideal es
1 2 2.63 3.13 3.56
Por tanto NDCG@5= 2.93/3.56 = 0.82

3. (1 punto) Considere un índice invertido donde se codifican en las listas invertidas los docID, tf y posiciones de los términos en los documentos. Los docID y posiciones se codifican con d-gaps y todo con variable byte encoding. Un término t tiene la siguiente lista invertida_
00000010 10000010 10000011 00000010 10000000 10000001 10000001 10000100 10000001 10000100
Indique en que documentos y posiciones aparece el término t, rellenando la siguiente tabla (la tabla puede necesitar más o menos filas de las que se muestran).

docID	pos1, pos2, etc.

Al codificar con variable byte encoding, si el primer bit (el bit mas significativo, es decir, el bit a la izquierda en el orden de lectura del byte) es 1, indica que es el último byte del número a codificar, mientras que si el primer bit es 0 indica que no es el último byte del número a codificar. Por tanto el bit mas significativo de cada byte es una marca y sólo se pueden usar los otro 7 bits para codificar los valores de los números. Por tanto al leer esa secuencia de bytes, los números en decimal que se corresponden son:
258 3 256 1 1 4 1 4
el primer documento es docID 258 con tf=3 y el término aparece en las posiciones 256, 257, 258
después aparece en el documento con docID 262 con tf=1 y aparece en la posición 4

4. a) (0.75) Escriba la expresión para computar P(D|Q) y P(Q|D) usando el modelo de recuperación de Language Models y suavización de Jelinek Mercer, explicando con claridad todos los elementos que aparecen en las expresiones.
b) (0.25) Imagine que los documentos son páginas web y se computó Page Rank. ¿Puede acomodar Page Rank de alguna manera en las expresiones anteriores? Si es posible diga como, en caso que no sea posible diga por qué.

5. (1 punto) Conteste V/F (Verdadero/Falso), o no conteste, a cada apartado de la pregunta. Cada apartado se puntúa con 0.1 puntos, pero **cada respuesta incorrecta invalida una correcta**. La pregunta no tendrá puntuación negativa. **Solo se evalúan las respuestas de la tabla y cualquier respuesta dudosa cuenta como incorrecta.**

1	2	3	4	5	6	7	8	9	10
V	F	F	F	V	V	F	V	F	F

1) Considere el hilo de crawling visto en clase. Con la implementación adecuada de la frontera puede hacerse crawling primero en anchura o primero en profundidad de los sitios webs.

2) El archivo robots.txt que especifica la política de exclusión de robots reside en el módulo del crawler encargado de cumplir con el requerimiento de politeness.

3) Considere el hilo de crawling visto en clase. Para adaptarlo para un hilo de un focused crawler, sería suficiente con usar como semillas páginas autoritarias de la temática con la que pretende tratar el seach engine.

4) Un índice invertido que almacena las posiciones de las palabras en los documentos, no podría almacenar también los byte-offsets de las palabras (el byte-offset indica el byte-offset del documento donde empieza la pabra, mientras la posición indica la posición de la palabra en el documento, es decir no contienen la misma información)

5) En un índice invertido, idf(t), inverso de la frecuencia de documento de un término t, no se almacena en el índice sino que se computa en tiempo de procesado de consultas.

6) El algoritmo de Rocchio fue ideado para la tarea de Real Relevance Feedback pero puede usarse en la tarea de Pseudo-Relevance Feedback.

7) En el modelo de Relevance Models, $P_{q1}(word|R)$ indica la probabilidad de una palabra en el modelo de relevancia para una query q1. Si un usuario formula una query q2, inmediatamente después de formular q1, y el search engine decide hacer Pseudo-Relevance Feedback, podría reusar $P_{q1}(word|R)$ como modelo de relevancia de q2.

8) Dada una query q, si un sistema de recuperación de información A tiene mejor P@5 que otro sistema B, entonces se garantiza que A también tiene mejor R@5 que B.

9) Dada una query q, si un sistema de recuperación de información A tiene mejor P@10 que otro sistema B, entonces se garantiza que A también tiene mejor AP@10 que B.

10) En un benchmark que tiene 100 queries de test, si un sistema de recuperación de información A tiene un MAP (Mean Average Precision) mayor del 10% que el MAP del sistema B, se garantiza que la diferencia es estadísticamente significativa con un nivel de significancia alpha=0.05, es decir, nivel del significancia del 5%, y medida con el t-test.