Examen **Recuperación de Información** Junio 2017

Apellidos: _____Nombre: _____

Examen sin libros, apuntes, ni dispositivos electrónicos.
Tiempo: 2h. 15m.

1. (0.5 puntos) Considere una query con 6 documentos relevantes, y un sistema de RI que devuelve un ranking con relevantes en las posiciones 2, 4, 5, 8, 14, 15 del ranking.

   a) (0.2) Calcule la precisión interpolada a los niveles estándar de recall

Recall	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Prec	0.6	0.6	0.6	0.6	0.6	0.6	0.5	0.4	0.4	0.4	0.4

El ranking sería a efectos de no relevantes y relevantes:
NRNRR NNRNN NNNRR

Podemos fijarnos en los pares (Recall, Precision) cada vez que aparece un relevante. Serían:
(1/6, ½), (2/6, ½) (3/6, 3/5), (4/6, ½), (5/6, 5/14), (6/6, 6/15) =
(0.17, 0.5), (0.33,0.5) (0.5, 0.6), (0.67,0.5), (0.83, 0.36), (1, 0.4) =

La precision interpolada para un valor de recall, es el mayor de los valores de precisión para todos los valores de recall a la derecha, es decir valores de recall mayor o igual que para el que se está obteniendo la precisión (slide 23 del capítulo de evaluación). Por tanto para todos los valores de recall hasta 0.5 incluido, la precisión interpolada será 0.6 y para los valores a la derecha será 0.4.

b) (0.2) Calcule el MAP con corte 20
MAP@20 = _____

El MAP ya lo calculé en la solución al examen de Julio 2017. Comprobad con vuestros compañeros que tenéis el mismo resultado.

c) (0.1) Calcule R-Precision

Sea R el número total de relevantes para una query. R-Precision es el valor de Precisión en el cutoff R del ranking. Esta query tiene 6 relevantes, entonces en este caso R-Precision = Precision@6= ½
Se cumple que en corte R del ranking, la precisión es igual al recall.

Observad también que para el ranking ideal R-Precision=1, lo que hace que el valor de esta métrica sea muy fácilmente interpretable porque nos dice lo que nos acercamos al ranking ideal.

2. (0.5 puntos) Considere el modelo booleano de RI y queries booleanas conjuntivas y considere los siguientes órdenes de procesamiento de las listas de postings en el índice invertido

a) orden alfabético creciente
b) orden alfabético decreciente
c) Frecuencia de documento (df(t)) creciente
d) Frecuencia de documento (df(t)) decreciente
e) Fecuencia de término (tf(t,d)) creciente
f) Fecuencia de término (tf(t,d)) decreciente

-¿Cuál es el mejor (o mejores) órdenes en términos de eficacia (effectiveness)?
Los órdenes de procesamiento de las listas invertidas dados por e) y f) no tiene sentido. Para los otros cuatro, la eficacia sería la misma. En todos los caso se devolverían los documentos que satisfacen la query conjuntiva.

-¿Cuál es el mejor (o mejores) órdenes en términos de eficiencia (efficiency)?

De nuevo e) y f) no tienen sentido. Lo más eficiente es empezar por la lista más pequeña (df(t) más pequeño) así el conjunto de resultados es mas pequeño, y al intersectar con las siguiente listas sabrás que a partir del último docID que tienes en el conjunto de resultados ya no tienes que seguir explorando esa lista. Por tanto el orden es el dado por c), es decir Frecuencia de documento (df(t)) creciente.

3. (0.75 puntos) Considere un grafo web con dos páginas d1 y d2. De d1 sale un enlace a d1, y de d2 sale un enlace a d2. Para este grafo web, ¿es posible calcular el Page Rank con un teleporting del 20%? Marque la respuesta correcta. Si es posible compútelo e indique los cálculos , en caso contrario diga por qué no se puede.

-Si.
Solución: _____

-No.
Razón: _____

Ya hice un ejercicio de Page Rank calculado matricialmente en el documento que dejé en moodle, otros en el examen de Julio 2017 y Julio 2018. Por favor hacer este ejercicio entre vosotros y comprobad con otros estudiantes el resultado.

4. (0.75 puntos) Suponga una colección de documentos que contiene dos documentos d1 y d2 con los contenidos que se muestran.
Contenido de d1: Tempus reports a profit but revenue is down
Contenido de d2: Quorus Global narrows quarter loss but revenue decreases further

Considere el modelo de RI Query Likelihood con MLE suavizado con Jelinek-Mercer con λ=0.5, y que el procesado de texto no hace stemming, ni filtrado de stop words, ni paso a minúsculas.
Considere la query q = revenue down
y compute en este modelo:

P(q | d1) = _____
P(q | d2) = _____

$P(q|d) = \Pi_i p(q_i | d) = \Pi_i ((1-\lambda) (f_{qi,d}/|d|) + \lambda (f_{qi,C}/|C|))$

P(q |d) es el query likelihood. Basta operar para d1 y d2 con los datos que resultan del enunciado:
|C| = 17, |d1|=8, |d2|=9
$f_{revenue,d1} = 1$

$f_{down,d1} = 1$

$f_{revenue,d2} = 1$

$f_{down,d2} = 0$

Considere ahora que se considera además el Prior (probabilidad a priori) de los documentos proporcional al número de palabras que empiezan con mayúscula. Vuelva a computar:

P(d1 | q) = _____
P(d2 | q) = _____

$P(d|q) = (P(q|d) P(d)/P(q)) = ^{rank} P(q|d) P(d)$

P(d1) = ⅓
P(d2) = ⅔
y los query likelihoods se computaron en la sección anterior.

5. (0.5) Considere un índice invertido donde las listas invertidas contienen sólo la información de los documentos donde ocurren los términos y se codifican con d-gaps.

a) ¿Cuál es el mayor d-gap que se puede codificar con 3 bytes usando variable-byte encoding?

Tenemos 21 bits para codificar en binario ya que un bit de cada byte es una marca
Por tanto $2^{21} -1$

b) El término *alfa* tiene la siguiente lista invertida:

00000011 10000011 10001001 11111110

¿En que números de documentos ocurre el término *alfa*? Es decir, escriba el docID de cada documento en que ocurre el término *alfa*.
Solución: ____387, __396, ____522_____

$2^8 + 2^7 + 2 + 1 = 387$

6. (de 0 a 2 puntos).
-Cada pregunta tipo test (1 al 10) vale 0.2 puntos pero **1, 2 preguntas incorrectamente contestada resta una correctamente contestada; 3, 4 preguntas incorrectamente contestadas restan dos correctas, etc.** Cada pregunta tiene una respuesta correcta. La contestación que se puntuará será la que aparezca en el espacio dedicado:
RESPUESTA :_____.

Contestaciones ambigüas se tomarán como incorrectas.

1. Al evaluar dos sistemas de RI A y B
RESPUESTA :_____

A. Para cualquier query, si A es mejor que B en P@10, también A es mejor que B en P@20
B. Para cualquier query, si A es mejor que B en Recall@10, también A es mejor que B en Recall@20
C. Para cualquier query, si A es mejor que B en AP@10, también A es mejor que B en AP@20. AP@k es AP computada hasta la posición k del ranking.
D. ninguna de los anteriores

2. Al hacer la comparación de dos sistemas *one sided* (A mejor que B) con el test t de significancia estadística y una métrica determinada
RESPUESTA:_____

A. Un p-valor de 0.03 indicaría que para esa métrica en promedio A es mejor que B en al menos un 3%
B. Que podemos rechazar la hipótesis nula si el nivel de significancia es mayor que el p-valor
C. Que podemos rechazar la hipótesis nula si el nivel de significancia es menor que el p-valor
D. ninguno de los anteriores

3. Cuales de los siguientes NO puede ser un index term
RESPUESTA:_____

A. un adverbio
B. un word stem (raíz de una palabra)
C. un n-gram de n caracteres
D. todos los anteriores SI pueden ser index terms

4. La ley de Heap y=f(x)
RESPUESTA:_____

A. relaciona la frecuencia de ocurrencia de una palabra (y) con su rango por frecuencia (x) en una colección de documentos
B. relaciona el tamaño del vocabulario (y) con el número de *index terms* (x) de la colección
C. relaciona el tamaño del vocabulario (y) con el número de documentos (x) de la colección
D. ninguna de las anteriores

5. En un servidor web el archivo robots.txt permite
RESPUESTA:_____

A. bloquear el acceso de ciertos crawlers
B. bloquear el acceso desde ciertas direcciones IP
C. especificar directorios a los que no debe acceder el crawler
D. ninguna de las anteriores.

6. Los sitemaps ayudan a los crawlers a realizar su tarea y son generados por
RESPUESTA:_____

A. un módulo del search engine de análisis de enlaces
B. un módulo del search engine de análisis de contenidos textuales
C. un módulo del search engine que analiza enlaces y contenidos textuales
D. ninguna de las anteriores.

7. Un algoritmo de detección de duplicados basado en técnicas de checksum
RESPUESTA:_____

A. sólo permite detectar duplicados exactos
B. permitiría detectar near-duplicates (cuasi-duplicados) pero sólo de imágenes o videos
C. permitiría detectar near-duplicates (cuasi-duplicados) pero sólo de textos
D. ninguno de los anteriores

8. En un índice en cuyas listas invertidas sólo se registran los documentos donde aparecen los términos con la técnica de d-gaps
RESPUESTA:_____

A. no puede computarse ningún modelo vectorial porque es un índice diseñado para el modelo booleano de IR
B. puede usarse para computar un modelo vectorial basado en similaridad de coseno y esquema de pesado *raw tf x idf* en queries y *tf binario x idf* en documentos
C. no puede computarse ningún modelo vectorial porque no se registra la frequencia de los términos ni en los documentos ni en las queries
D. ninguna de las anteriores

9. Los search engines
RESPUESTA:_____

A. no pueden hacer caching de listas de resultados porque éstas son inmensas
B. no pueden hacer caching de listas de los índices invertidos porque éstas son inmensas
C. no pueden hacer caching de resultados ni de listas de los índices invertidos porque cambian muy dinámicamente
D. hacen caching de listas de resultados comunes y listas de los índices invertidos de términos frecuentes

10. Un problema con el modelo booleano de RI es que
RESPUESTA:_____
A. no se pueden procesar consultas sobre índices invertidos
B. para procesar consultas es necesario el direct file (índice directo)
C. no puede ser un modelo eficaz de RI
D. ninguna de las anteriores