

Examen **Recuperación de Información** Julio 2021

Apellidos: _____ Nombre: _____

Examen sin libros, apuntes, ni dispositivos electrónicos.

Tiempo: 2h. 15m.

1. (1 punto) Considere un grafo web con 5 nodos (1, 2, 3, 4, 5) 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. Del nodo 5 sale un enlace a 2 y otro a 4. Compute la matriz de transición de probabilidad con un teleporting del 30% (MTP30%) y el Page Rank.

Page Rank: 0.06 0.41 0.06 0.41 0.06

(MTP30%) :

**0.06 0.41 0.06 0.41 0.06
0.06 0.76 0.06 0.06 0.06
0.06 0.41 0.06 0.41 0.06
0.06 0.06 0.06 0.76 0.06
0.06 0.41 0.06 0.41 0.06**

2. a) (0.5 puntos) El pseudocódigo de inversión en memoria **que se adjunta** tiene dos limitaciones importantes. Indíquelas.

b) (0.5 puntos) Indique las otras soluciones estudiadas en la asignatura que abordan estas limitaciones.

Use el anverso de esta página para contestar a) y el reverso para contestar b). Las respuestas deben ser claras y contenidas en ese espacio.

Fue explicado en clase y está disponible en las slides, libro y grabación de la clase.

3. (1 punto) **Se adjunta** el pseudocódigo del hilo de crawler estudiado en la asignatura. ¿Puede adaptarse y/o extenderse para construir un hilo de un focused crawler? En caso negativo, explique la razón de la imposibilidad. En caso afirmativo reescriba el nuevo hilo completo y explique la adaptación y/o extensión.

Fue explicado en clase y está disponible en las slides, libro y grabación de la clase.

4. (0.75 puntos) Suponga una colección de documentos que contiene 10 documentos. Con los caracteres a, b, c, etc. nos referimos a *index terms*. Los documentos d1 y d2 tienen los contenidos que se muestran. Los documentos d3, d4, d5, d6 son una copia de d1 y los documentos d7, d8, d9, d10 son una copia de d2

Contenido de d1: a b d f a b

Contenido de d2: b c d g b c

Considere el modelo de RI Query Likelihood con MLE suavizado con Jelinek-Mercer con $\lambda=0.5$

Considere la query $q = a c$

y compute en este modelo:

$$P(q | d2) = \underline{\hspace{2cm}}$$

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

$P(q | d)$ es el query likelihood. Basta operar para d2 con los datos que resultan del enunciado:

$$|C| = 60, |d2| = 6$$

$$f_{a, d2} = 0$$

$$f_{c, d2} = 2$$

$$f_{a, C} = 10$$

$$f_{c, C} = 10$$

$$P(q | d) = (0.5 \times 0 + 0.5 \times 1/6) \times (0.5 \times 1/3 + 0.5 \times 1/6) = 0.021$$

(0.25 puntos) Considere ahora que se considera además el Prior (probabilidad a priori) de los documentos uniforme.

Ignorando $P(q)$ (el Prior de la query), compute:

$$P(d2 | q) = \underline{\hspace{2cm}}$$

$$P(d | q) = (P(q | d) P(d) / P(q)) =_{\text{rank}} P(q | d) P(d)$$

$$P(d2) = 1/10$$

$$P(d2 | q) =_{\text{rank}} 0.021 \times 1/10 = 0.0021$$

5. (1 punto). En cada apartado (0.25 puntos) debe contestar con precisión y claridad lo que se pregunta en el espacio reservado.

- Para un benchmark de 50 queries como se computa [Recall@10](#)

-Para una query como se computa [AP@5](#) (AP es Average Precision)

-¿Qué es y cómo es un TREC Topic?

-¿Qué es y para qué es usa el proceso de Pooling en TREC?

Los contenidos de esta pregunta fueron explicados en clase y está disponible en las slides, libro y grabación de la clase.