Recuperación de Información

Diciembre 2014

1. Dado el siguiente índice posicional (cada posting indica doc ID seguido por posiciones):

   descr $\rightarrow <2(36,174,252,651)><4(12,22,102,432)><7(3,17)>$

   breve $\rightarrow <2(1,17,74,222)><5(8,78,108,458)><7(3,13,23,193)>$

   pasos $\rightarrow <3(87,704)><4(13,43,113)><7(18,328,528)><9(20,320)>$

   proce $\rightarrow <2(3,37,76,444,851)><5(10,20,110,470,500)><7(5,15,25,195)>$

   tipic $\rightarrow <2(2,66,194,253,702)><4(9,69,149,429,569)><7(4,14,404)>$

   index $\rightarrow <3(47,86,234,999)><7(14,24,774,944)><9(199,319,599,709)>$

   siste $\rightarrow <2(57,94,333)><4(15,35,155)><7(20,320)>$

   recup $\rightarrow <2(67,124,393,1001)><6(11,41,101,421,431)><9(16,36,736)>$

   ¿Qué documentos satisfacen las consultas a) `pasos recup` b) `descr tipic` c) `\"descr tipic"`

2. Describe el proceso de indexación basada en bloques

3. Supóngase que tenemos una colección de ~~cuatro~~ ocho documentos descrita por la siguiente matriz término/documento, donde las celdas indican el número de apariciones de los términos en cada documento.

	$d1$	$d2$	$d3$	$d4$	$d5$	$d6$	$d7$	$d8$
descr	-	-	-	-	4	1	4	2
breve	1	4	-	2	-	-	-	1
pasos	2	-	2	-	4	1	-	-
proce	-	-	-	-	8	-	-	-
siste	-	4	-	-	4	1	-	8
recup	-	4	-	-	-	-	4	-

   a) Calcular el peso tf-idf de los términos en cada documento.

   b) Utilizando la similitud por coseno, calcular el valor de recuperación (score) de los documentos $d8$ y $d4$ para la consulta: `descr`*ipción* de un `sistem`*a*.

4. ¿Cómo calcula el score el Language Model (Modelado del Lenguaje)?. Pon un ejemplo (considerando los datos del Ejercicio3 con valor $\lambda = 0{,}3$) para el documento $d5$ sobre la consulta: `descr`*ipción* de un `sistem`*a*.

5. Para una consulta se han recuperado la siguiente secuencia de documentos

   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   N R R N N R R N N R N N R N N R N N N N
   (R en posición 10)

   Suponiendo que en la colección hay un total de 10 documentos relevantes, calcular la exhaustividad, la precisión, P@10, AP@20 y NDCG@5.

Notas

Tiempo para realizar el examen: 2 horas.

$= \frac{DCG@5}{DCGideal(hasta\ 5)}$

$\frac{\frac{1}{2}+\frac{2}{3}+\frac{3}{6}+\frac{4}{7}+\frac{5}{10}+\frac{6}{13}+\frac{7}{16}}{7}$