ACTO2 – SAR

(22/05/2017 - 3 puntos)

(IMPORTANTE: Se deben justificar las respuestas)

1) Se pide dar la secuencia de bis correspondientes a la compresión por códigos gamma de la siguiente posting list: [7, 12, 13, 16, 20, 26, 34, 35, 43]: **(0,5 puntos)**

Solución:

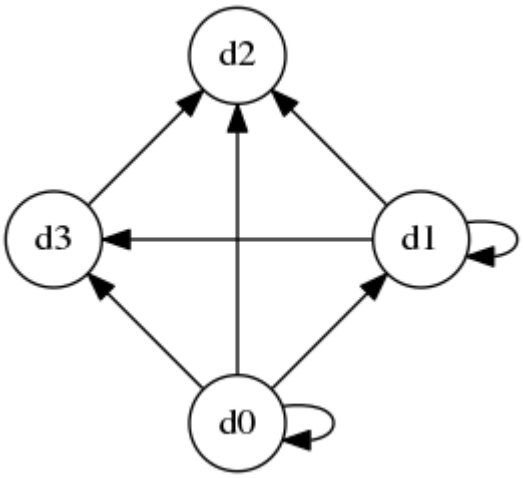
La secuencia de gaps en decimal es: [7, 5, 1, 3, 4, 6, 8, 1, 8]
La correspondiente secuencia de bits es:
11011 11001 0 101 11000 11010 1110000 0 1110000

2) Esta pregunta consta de dos apartados: **(0,5 puntos)**
   a. Enuncia la ley de Heap y justifica su utilidad en recuperación de información.
   b. Disponemos de dos colecciones diferentes de documentos que cumplen la ley de Heap. Suponiendo que ambas colecciones tienen el mismo tamaño en cuanto a número de tokens, ¿tendrán exactamente el mismo tamaño de vocabulario?

Solución:

a. La ley de Heap dice $M=kT^b$, donde K y b son parámetros que dependen de la colección. En RI se utiliza para hacer una estimación del tamaño del vocabulario de la colección de documentos. Esta información es importante cuando se aborda la implementación del índice invertido.
b. No
   La ley de Heap dice $M=kT^b$, donde K y b son parámetros que dependen de la colección. Por tanto, aunque las dos colecciones tengan el mismo valor para el parámetro T, el resultado de aplicar la ley puede proporcionar valores diferentes para M, dependiendo de los valores de los parámetros K y b de cada colección.

3) Dadas las siguientes páginas web y los enlaces entre ellas representadas como un grafo, se pide calcular el pagerank de cada página. Se debe calcular: i) la matriz de enlaces, ii) la matriz de probabilidades de transición, iii) la matriz de probabilidades de transición con teletransporte (utiliza un α=0,1 para el teletransporte), iv) todas las iteraciones para calcular el pagerank. Realiza como máximo cinco iteraciones.
(1 punto)



Solución:

Matriz de enlaces
[[1 1 1 1]
 [0 1 1 1]
 [0 0 0 0]
 [0 0 1 0]]

Matriz de probabilidades de transición
[[0.2500 0.2500 0.2500 0.2500]
 [0.0000 0.3333 0.3333 0.3333]
 [0.0000 0.0000 0.0000 0.0000]
 [0.0000 0.0000 1.0000 0.0000]]

Matriz de probabilidades de transición con teletransporte
[[0.2500 0.2500 0.2500 0.2500]
 [0.0250 0.3250 0.3250 0.3250]
 [0.2500 0.2500 0.2500 0.2500]
 [0.0250 0.0250 0.9250 0.0250]]

Iteraciones cálculo de PageRank
$\overrightarrow{x_0}$ = (1.0000, 0.0000, 0.0000, 0.0000)
$\overrightarrow{x_1}$ = (0.2500, 0.2500, 0.2500, 0.2500)
$\overrightarrow{x_2}$ = (0.1375, 0.2125, 0.4375, 0.2125)
$\overrightarrow{x_3}$ = (0.1544, 0.2181, 0.4094, 0.2181)
$\overrightarrow{x_4}$ = (0.1519, 0.2173, 0.4136, 0.2173)
$\overrightarrow{x_5}$ = (0.1522, 0.2174, 0.4130, 0.2174)
$\vec{\pi}$ = (0.1522, 0.2174, 0.4131, 0.2174)

4) Se quiere buscar el patrón “COSO” en el texto “TOSOESPUCOSAS”. Para ello se hace una búsqueda aproximada para obtener aquellos segmentos cuya distancia al patrón es menor o igual que 1. Se pide construir la matriz que corresponde al algoritmo de búsqueda aproximada e indicar las soluciones, es decir, los segmentos del texto que son resultados de la búsqueda. Se consideran pesos 1 para las Sustituciones, Inserciones y Borrados. **(1 punto)**

O	4	4	3	2	**1**	2	3	3	4	3	2	**1**	**1**	2
S	3	3	2	1	2	2	2	3	3	2	1	0	1	2
O	2	2	1	2	1	2	2	2	2	1	0	1	2	2
C	1	1	1	1	1	1	1	1	1	0	1	1	1	1
	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		T	**O**	**S**	**O**	**E**	**S**	**P**	**U**	**C**	**O**	**S**	**A**	**S**

Solución: TOSO: 1-4, OSO:2-4, COS: 9-11, COSA:9-12