ACTO2 – SAR

(18/06/2018 - 3 puntos)

Apellidos y Nombre:……

(IMPORTANTE: se pide justificar las respuestas)

1) Se pide obtener la postings list a partir de la siguiente secuencia de bits codificada utilizando codificación variable en bytes: **(0,5 puntos)**

00000100 10000001 10000111 00000001 00010100 10001101

Solución:

La decodificación de la secuencia anterior siguiendo el esquema de codificación variable en bytes corresponde a la secuencia de gaps [513, 7, 18957], por lo que la postings list es [513, 520, 19477].

2) Esta pregunta consta de dos apartados: **(1 punto)**

a) Enuncia la ley de Zipf y justifica su utilidad en recuperación de información.

b) Asumiendo que la longitud de la postings list de una colección de documentos sigue una ley de Zipf ~ i^{-1}, y que las primeras 100 listas más largas tienen una longitud >= 1000, ¿qué posición ocupan en el ranking de frecuencias los términos que ocurren una vez?

Solución:

a) En lenguaje natural, hay unos pocos términos muy frecuentes y muchos términos que aparecen con baja frecuencia. Ley de Zipf establece que el i-ésimo término más frecuente tiene una frecuencia proporcional a 1/i. Siendo K una constante y cf_i la frecuencia que corresponde al término que ocupa la posición i en el ranking de frecuencias de términos, se enuncia la Ley de Zipf como: $cf_i = K/i$.
Esta ley empírica es útil para hacer una estimación de las longitudes de las postings list en un índice inertido.

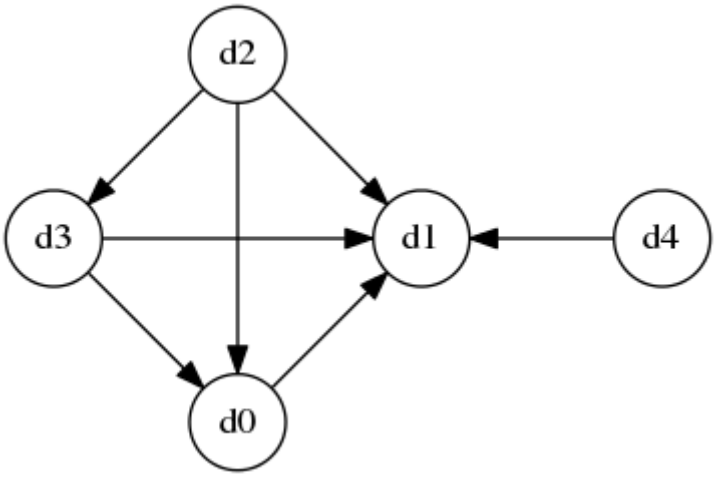
b) Si aplicamos la ley para i=100 y cf_i =1000, nos devuelve un valor de K=100x1000= 10^5
Ahora la aplicamos para el caso de una frecuencia 1, $1=10^5/i$, por lo que $i=10^5/1$.
Por tanto, los términos de frecuencia 1 ocupan una posición 10^5 en el ranking de frecuencias para este caso.

3) Se pide indicar sobre la tabla, los desplazamientos que se realizarían en una búsqueda por Booyer-Moore del patrón " **DBDDBF** " en la cadena " **FBABBFDADAWDBDDBFDCCA** ". **(0,5 puntos)**

Solución:

F	**B**	**A**	**B**	**B**	**F**	**D**	**A**	**D**	**A**	**W**	**D**	**B**	**D**	**D**	**B**	**F**	**D**	**C**	**C**	**A**
D	**B**	**D**	**D**	**B**	**F**															
	D	**B**	**D**	**D**	**B**	**F**														
			D	**B**	**D**	**D**	**B**	**F**												
					D	**B**	**D**	**D**	**B**	**F**										
											D	**B**	**D**	**D**	**B**	**F**				

4) Dadas las siguientes páginas web y los enlaces entre ellas representadas como un grafo, se pide calcular los valores HUB y AUTHORITY de cada página utilizando la aproximación HITS. Realiza cinco iteraciones sin normalización. **(1 punto)**



Solución:

Matriz de enlaces:

[0 1 0 0 0]
[0 0 0 0 0]
[1 1 0 1 0]
[1 1 0 0 0]
[0 1 0 0 0]

	HUBS	AUTHORITY
t_0	[1 1 1 1 1]	[1 1 1 1 1]
t_1	[1 0 3 2 1]	[2 4 0 1 0]
t_2	[4 0 7 6 4]	[5 7 0 3 0]
t_3	[7 0 15 12 7]	[13 21 0 7 0]
t_4	[21 0 41 34 21]	[27 41 0 15 0]
t_5	[41 0 83 68 41]	[75 117 0 41 0]

Hubs: [41 0 83 68 41]
Authority: [75 117 0 41 0]