Recuperación Acto2 - SAR

(17/06/2016 - 3 puntos)

Apellidos y Nombre:…….

(NOTA: Se pide justificar las respuestas. Los cálculos se mostrarán redondeados a 3 decimales)

1) Disponemos de una colección con un total de 10^7 documentos con una media de 12.000 caracteres por documento y una media de 8 caracteres de longitud por palabra. Suponiendo que esta colección satisface la ley de Heap con un valor b=0,5 y K=40, se pide estimar el número de palabras diferentes en la colección. **(0,5 puntos)**

Solución:

La ley de Heap dice M=kTb, por lo que necesitamos conocer el número de palabras/tokens de la colección T.
Como la media de caracteres por documento es 12.000 y la media de caracteres por palabra es 8, obtenemos una media de 1.500 palabras por documento. Por tanto el número total de palabras en los 10^7 documentos es de $1,5 \times 10^{10}$.
Aplicando la ley de Heap obtenemos un valor para M de $40 \times (1,5 \times 10^{10})^{0.5}$, que devuelve un resultado de 4.989.979,485 términos.

2) Se pide obtener la postings list a partir de la siguiente secuencia de bits codificada utilizando códigos gamma: **(0,5 puntos)**

1110001 11011 101 0 1110011 1110010 111101000 11010

Solución:

La secuencia de gaps en decimal es: 9, 7, 3, 1, 11, 10, 24, 6

La correspondiente posting list es:

[9, 16, 19, 20, 31, 41, 65, 71]

3) Se pide dar la secuencia de bits correspondientes a la compresión utilizando codificación variable en bytes de la siguiente postings list: **(0,5 puntos)**

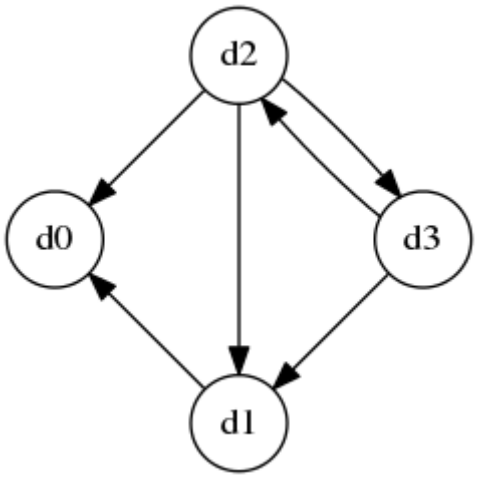
[788, 798, 19755]

Solución:
La secuencia de gaps en decimal es: [788, 10, 18957].
La correspondiente secuencia de bits utilizando codificación variable en bytes es:
00000110 10010100 10001010 00000001 00010100 10001101
Es decir: (6, 20+128) (10) (1, 20, 13+128)

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

F	E	D	E	C	E	D	B	W	E	C	F	A	A	F	C	E	D	D	D	W			
F	A	A	F	C	E																		
	F	A	A	F	C	E																	
							F	A	A	F	C	E											
										F	A	A	F	C	E								
											F	A	A	F	C	E							

5) 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 0 0 0]
[1 0 0 0]
[1 1 0 1]
[0 1 1 0]

	HUBS	AUTHORITY
t_0	[1 1 1 1]	[1 1 1 1]
t_1	[0 1 3 2]	[2 2 1 1]
t_2	[0 2 5 3]	[4 5 2 3]
t_3	[0 4 12 7]	[7 8 3 5]
t_4	[0 7 20 11]	[16 19 7 12]
t_5	[0 16 47 26]	[27 31 11 20]

Hubs: [0, 16, 47, 26]
Authority: [27, 31, 11, 20]