

# 7506-2020-2 Parcialito de IR

Thiago Kovnat

TOTAL POINTS

**100 / 100**

QUESTION 1

**1 IR 100 / 100**

✓ - **0 pts** Correct

Thiago Kovnat - 104429

1) Tengo el siguiente archivo de lexico concatenado:

Amarillentooncestralaquelrbolustoazul

0

y los siguientes punteros a documentados concatenados en codigo gamma:

010010 011 00100 1 010010 1010 10101

Noto que tengo un front coding parcial por lo que en cada acceso deberia armar la palabra basandome en las anteriores si es que no estoy en un indice que tenga 0 en comun con la palabra anterior.

Resuelvo palabra a palabra:

Aquel:

Procedo a hacer busqueda binaria sobre el indice, accedo a la posicion del medio, que me lleva al lexico en la posicion 20 (1 acceso a disco). Tengo 0 letras iguales a la anterior y 5 distintas. La palabra que accedi es aquel, que es la que estoy buscando. Ahora busco en los punteros a documentos la informacion de los mismos (1 acceso a disco). Me manda a la posicion 14 en la cual tengo unicamente un 1, lo cual me indica que esta en el documento 1.

Arbusto:

Accedo a la mitad de la tabla, la cual sabemos que es la palabra aquel por lo tanto me quedo con la mitad inferior viendolo en el esquema presentado en el parcial. Accedo a la mitad de la tabla, lo cual me manda a la posicion 29 del lexico (1 acceso a disco), pero tengo que usar la palabra anterior para formar la actual, por lo que tengo (1 acceso a disco mas). La palabra en la posicion anterior tambien necesita de la previa para formarse, pero esa posicion ya la habiamos leído y era aquel, por lo que la palabra previa a la que quiero analizar es arbol y con esa informacion puedo formar mi palabra actual ARBUSTO la cual es la que estaba buscando. El puntero en esa posicion me manda a la posicion 21 (1 acceso a disco). Tengo 1010. El 1 se refiere al documento 1 y el 010 al documento 3, ya que es un 2 y  $1+2=3$

Azul:

Accedo a la mitad de la tabla, otra vez mi palabra es aquel por lo cual me quedo con la parte inferior al igual que en el anterior. Vuelvo a acceder a la mitad que se que es Arbusto, por lo cual solo me queda una posibilidad que es acceder a la ultima posicion que me manda a la posicion 33 del lexico (1 acceso a disco). Mi palabra es azul que es la que estoy buscando. Veo en los punteros a docs a donde tengo que ir, que es la posicion 25 (1 acceso a disco). En la pos 25 tengo 10101. El 1 hace referencia al documento 1, el 010 al documento 3 ya que  $1+2=3$  y el 1 final hace referencia al documento 4 ya que  $3+1=4$ .

En total, tuve 7 accesos a disco.

Por lo tanto, tengo la siguiente informacion:

Aquel: Doc 1

Arbusto: Doc 1,3

Azul: Doc 1,3,4

Para saber que documentos contienen las 3 palabras debo hacer la interseccion de

estos datos, por lo que el documento que queda es el 1.

b) Como no tengo la frecuencia de cada termino en los documentos, voy a asumir que solo aparecen 1 vez.

Debo calcular el IDF para cada termino:

Se que el IDF se calcula como  $\log(N+1/ft_i)$

Ademas, habria que tener en cuenta que en el punto anterior, por mas que no se hizo referencia a un documento 2, este existe.

Calculo el IDF para cada termino:

Aquel:  $\log(5/1) = \log(5) = 0.699$  (aprox)

Arbusto:  $\log(5/2) = 0.398$  (aprox)

Azul:  $\log(5/3) = 0.222$  (aprox)

Ahora puedo rankear cada documento:

Para el doc 1:  $1 * 0.699 + 1 * 0.398 + 1 * 0.222 = 1.319$

Para el doc 3:  $1 * 0.398 + 1 * 0.222 = 0.62$

para el doc 4:  $1 * 0.222 = 0.222$

Por lo tanto, el orden de los documentos quedaria:

- 1) Doc 1
- 2) Doc 3
- 3) Doc 4

1 IR 100 / 100

✓ - 0 pts Correct