# PARCIALITO INFO RETRIEVAL


## Franco Nicolás Batastini – 103775


### Information Retrieval 2020 2C 1 Enunciado

Tenemos un índice invertido con la siguiente estructura:

| = | >< | Lexico | Docs |
|---|---|---|---|
| 0 | 11 | 0 | 0 |
| 7 | 1 | 11 | 6 |
| 1 | 8 | 12 | 9 |
| 0 | 5 | 20 | 14 |
| 1 | 4 | 25 | 15 |
| 3 | 4 | 29 | 21 |
| 0 | 4 | 33 | 25 |

El siguiente archivo de léxico concatenado:

Amarillentooncestralaquelrbolustoazul

Y los siguientes punteros a documento concatenados, codificados con códigos gamma:

01001001100100101001010101010101

a) Resolver la consulta “aquel arbusto azul” indicando detalladamente cada paso. Informar la cantidad de accesos a disco necesarios para la resolución de la misma.

b) Para la consulta rankeada "aquel arbusto azul" determinar el TF.IDF de cada documento resultado de la búsqueda e indicar cómo quedaría el orden de los documentos resultado de la misma.

## RESOLUCION

a) Para empezar, primero obtendré los términos que se encuentran almacenados con front coding parcial de n=3. Lo cual me resulta en que los términos presentes en el documento son:

Amarillento o ncestral aquel rbol usto azul

- Amarillento
- Amarillo
- Ancestral
- Aquel
- Árbol
- Arbusto
- Azul

Entonces, conociendo los términos, ahora puedo decodificar los códigos gamma de los documentos, considerando que estos están almacenados como distancias respecto del anterior:

$$010\ 010_{6}011_{9}00100_{14}1_{15}010\ 010_{21}1\ 010_{25}1\ 010\ 1$$

La lista de documentos para los términos queda:

- Amarillento: D2, D4
- Amarillo: D3
- Ancestral: D4
- Aquel: D1
- Árbol: D2, D4
- Arbusto: D1, D3
- Azul: D1, D3, D4

Entonces, ahora teniendo tanto los términos, como en que documentos están presentes, puedo resolver la consulta. De esta forma, para buscar a cada uno de los términos se debe realizar una búsqueda binaria en los términos:

*Busco a Aquel*

i. Ingreso al índice en búsqueda del término del centro. Será el de la posición 2 (Tercer término), que es Ancestral. Aquel es mayor que ancestral, por ende sigo con búsqueda binaria para los términos de más adelante. (Accesos a: Índice para saber dónde está el léxico, índice siguiente para saber dónde termina, al léxico del término, al léxico anterior, y al anterior del anterior a este para saber que letras comparte. Accesos totales = 5 a disco).

ii. Ingreso al índice, el término obtenido es Árbol, que comparte una letra con el término anterior a este. No es, por lo tanto aplico de nuevo búsqueda binaria (Accesos: Índice 2 veces para saber dónde empieza y termina, al término anterior para saber que letra comparte, y al término en sí. Accesos totales = 4 a disco).

iii. Aquel es menor que árbol, por lo tanto se accede con búsqueda binaria del lado izquierdo. Esta vez si encontramos a aquél, y tomamos los documentos de este (Accesos: Los de índice ya los tengo en memoria por el paso 2 (2 accesos cacheados), el acceso al término también estará cacheado (1 acceso cacheado), no comparte letras con otros términos, entonces finalizan los accesos; resta obtener los documentos, por lo que habrá dos accesos al índice para conocer las delimitaciones en el código gamma, y 1 acceso más al código gamma para traerme los documentos (3 accesos). Accesos totales = 3 accesos cacheados + 3 accesos a disco).

Finalmente los documentos obtenidos para aquél son: D1.

*Busco a Arbusto*

i. Mismo paso que en búsqueda de Aquel, solo que ahora los 5 accesos estarán cacheados (5 accesos cacheados).
ii. Arbusto es mayor que Ancestral, por lo tanto analizamos el lado derecho. Mismo paso que en Aquel. (habrá 4 accesos cacheados).
iii. Arbusto es mayor que árbol, entonces se busca en el lado derecho. Encontramos a arbusto, habrá dos accesos al índice para conocer los límites del léxico, dos más para conocer el límite de los documentos (4 accesos a disco), se tendrá que acceder a los términos árbol y aquel que se encuentran cacheados anteriormente para conocer los caracteres que arbusto comparte (2 accesos cacheados), y al término de arbusto en sí(1 acceso a disco); finalmente se accede a disco para obtener los documentos de árbol (1 acceso a disco) (Accesos totales = 6 accesos a disco + 2 accesos cacheados).

Finalmente los documentos obtenidos para arbusto son: D1, D3

*Busco a Azul*

i. Se repetirían todos los pasos de arbusto con accesos cacheados (17 accesos cacheados).
ii. Se realiza el último paso de la búsqueda, obtendríamos al término azul, con 2 accesos al índice en disco (1 para el inicio del término, y 1 para el inicio de los documentos); no son 4 porque se encuentra el fin de nuestra estructura. No se accede a ningún término adicional, pues no comparte caracteres. Se accede a disco para obtener el término azul, y otro acceso más para obtener los documentos de este (2 accesos a disco). (Accesos totales = 4 accesos a disco).

Resultado final para azul: D1, D3, D4

El resultado final se obtiene operando con un AND entre los documentos que contienen a Azul, a Arbusto, y a Aquel.

Aquel AND Azul AND Arbusto =

= D1 AND D1,D3,D4 AND D1,D3 =

= D1 AND D1,D3 =

= D1

- Resultado que contenga todos los términos: D1.
- Accesos totales a disco: 22
- Accesos totales cacheados: 31

b) Asumiendo la frecuencia de cada término como 1 en cada uno de los documentos, realizo la tabla correspondiente a TF-IDF:

| | TF | | | | IDF |
|---|---|---|---|---|---|
| | D1 | D2 | D3 | D4 | Log(N+1/tfi) |
| Amarillento | - | 1 | - | 1 | 0,39794 |
| Amarillo | - | - | 1 | - | 0,69897 |
| Ancestral | - | - | - | 1 | 0,69897 |
| Aquel | 1 | - | - | - | 0,69897 |
| Árbol | - | 1 | - | 1 | 0,39794 |
| Arbusto | 1 | - | 1 | - | 0,39794 |
| Azul | 1 | - | 1 | 1 | 0,22185 |

Siendo N la cantidad de documentos, y tfi, la cantidad de documentos que contienen al término i.

Log (5/1) = 0,69897
Log (5/2) = 0,39794
Log (5/3) = 0.22185

Finalmente para ranquear la consulta, se toman los documentos que poseen a los términos del query; en este caso “aquel arbusto azul”, y se multiplica el tf por el idf, y se los suma para cada uno de los términos listados en el query.
Los documentos que no contengan a ninguno de los términos tendrán un Rank de 0, y no se considerarían como resultado de la búsqueda.

- D1: 1 . 0,69897 + 1 . 0,39794 +1 . 0,22185 = 1,31876
- D2: 0
- D3: 1. 0,39794 + 1. 0,22185 = 0,61979
- D4: 1 . 0,22185 = 0,22185

| RANKING DE CONSULTA “AQUEL ARBUSTO AZUL” |
|---|
| D1 = 1,31876 |
| D3 = 0,61979 |
| D4 = 0,22185 |
| D2 = 0 |