5) Dada la siguiente colección de documentos:

D1: AZUL MARRON AZUL ROSA
D2: VERDE AZUL AMARILLO ROJO BLANCO
D3: VERDE NEGRO BLANCO
D4: BLANCO NEGRO
D5: VERDE VERDE NEGRO VERDE

Construya un índice invertido, indicando el paso a paso en la construcción del mismo y seleccionando un método de almacenamiento de punteros y del léxico que considere adecuado para este caso, justificando su elección.
En el índice construído, resuelva la consulta por frase "VERDE NEGRO", explicando el paso a paso en su resolución y contabilice la cantidad de accesos necesarios para su resolución.

(***) (15 pts)

Pasos:

1. Debemos recorrer secuencialmente la colección de archivos armando un archivo auxiliar ordenado alfabéticamente en donde indiquemos el término y su documento, la frecuencia y sus posiciones. Aclaramos que utilizamos distancias.

Amarillo: 2; 1 (3)
Azul: 1; 2 (1, 2), 1; 1 (2)
Blanco: 2; 1 (5), 1; 1 (3), 1; 1 (1)
Marrón: 1; 1 (2)
Negro: 3; 1 (2), 1; 1 (2), 1; 1 (3)
Rojo: 2; 1 (4)
Rosa: 1; 1 (4)
Verde: 2; 1 (1), 1; 1 (1), 2; 3 (1,1,2)

2. Generamos el listado de términos concatenados, utilizando la concatenación de términos. Nos quedaría:

0 8 13 19 25 30 34 38
amarillo azul blanco marron negro rojo rosa verde

3. Ahora vamos armar el listado de punteros (distancias), para eso vamos a utilizar la codificación gamma.

gamma(1) = 1
gamma(2) = 010
gamma(3) = 011
gamma(4) = 00100

0 7 20 37
010 1 011 | 1 010 1,010 , 1 1 010 | 010 1 00101 , 1 1 011 , 1 1 1 | 1 1 010
42 59 68 75
| 011 1 010 , 1 1 010 , 1 1 011 | 010 1 00100 | 1 1 00100 | 010 1 1 , 1 1 1 , 010 011 1,1,010

Entonces la pablovich queda:

Términos	Léxico	Docs
Amarillo	0	0
azul	8	7
blanco	13	20
marrón	19	37
negro	25	42
rojo	30	59
rosa	34	68
verde	38	75

Resolviendo la consulta: “Verde Negro”

- Se debe resolver las consultas puntuales para uno de los términos.
- Se obtiene la lista de documentos donde aparecen todos los términos.
- Para cada documento se validan las posiciones de cada término.

Realizamos la consulta de verde:

pos 3 (marrón) (1 índice, 1 disco) leemos 6 chars
pos 5 (rojo) (1 índice, 1 disco) leemos 4 chars
pos 6 (rosa) (1 índice, 1 disco) leemos 4 chars
pos 7 (verde) (1 índice, 1 disco) leemos 5 chars

Habiendo encontrado la palabra buscada, leemos los punteros a documentos (1 disco), desde la posición 75 hasta que termine.

010 1 1 , 1 1 1 , 010 011 1,1,010

Decodificando concluimos que el término verde aparece en los documentos 2, 3, y 5.
Doc2: pos 1.
Doc3: pos 1.
Doc5: pos 1, 2 y 4.

Realizamos la consulta de negro:

pos 3 (marrón) Ya lo leímos.
pos 5 (rojo) Ya lo leímos.
pos 4 (negro) (1 indice, 1 disco) leemos 5 chars

Habiendo encontrado la palabra buscada, leemos los punteros a documentos (1 disco), desde la pos 42 a 59.

011 1 010 , 1 1 010 , 1 1 011

Decodificando obtenemos que el término aparece en los documentos 2, 3 y 4.
Doc2: pos 2
Doc3: pos 2
Doc4: pos 3

Ya tenemos las consultas puntuales de ambos términos, por lo tanto, procedemos a descartar documentos, nos quedamos solo con los que aparecen todos los términos. Entonces nos quedamos solo con los documentos 2 y 3.

Ahora, pasamos a validar las posiciones en los distintos documentos

Doc2: verde (pos 1) negro (pos 2)
Posee la frase.

Doc3: verde (pos 1) negro (pos 2)
Posee la frase.

El resultado de la búsqueda de frase son los documentos 2 y 3.