Santiago Locatelli - 104107
Parcialito Information Retrieval

A)

Primero debemos recorrer secuencialmente la colección de documentos armando un archivo auxiliar en donde indiquemos el término y su documento. Este archivo auxiliar debe ser ordenado por término y documento.

Entonces quedaría:

Términos	Documentos
cabalgar	2, 4
Caballero	1, 4
caballo	1, 2, 4
cabaña	3
cabra	1, 3, 5
cacao	2, 3
cacerola	2, 5

Ahora recorremos el archivo secuencialmente para generar el listado de términos concatenados, los generamos con front coding parcial de n=3. Nos queda:

=	!=	char
0	8	0
5	4	8
6	1	12
0	6	13
3	2	19
2	3	21
0	8	24

Santiago Locatelli - 104107
Parcialito Information Retrieval

0 8 12 13 19 21 24
cabalgarleroo cabañaracao cacerola

Lo próximo que sigue es pasar las referencias a documentos a distancia, y codificar las distancias (el enunciado no especifica nada sobre usar distancias, pero siempre las codificaciones se ven beneficiadas si se usan números más pequeños, entonces decidí usar distancia). Y generar el listado de distancias concatenadas.

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

0 6 10 15 18 26 30
010010 1011 11010 011 1010010 0101 010011

Por lo tanto la estructura del índice quedaría de la siguiente forma:

=	!=	char	doc
0	8	0	0
5	4	8	6
6	1	12	10
0	6	13	15
3	2	19	18
2	3	21	26
0	8	24	30

B)

Buscamos “caballo”

Procedemos a hacer la búsqueda binaria del término. Accedemos a la mitad del índice, sería la posición 3, como vemos que los caracteres iguales son 0, entonces podemos leer todos los caracteres desde la posición 13 hasta la 18. Vemos que no es la palabra que buscamos, como caballo es menor alfabéticamente nos quedamos con las posiciones inferiores.

Entonces accedemos a la posición 1 del índice, notamos que tenemos que rearmar la palabra, ya que la cantidad de caracteres iguales es 5, accedemos a la posición 0, leemos desde la posición 0 hasta la 4 del listado de términos concatenados, y leyendo desde la pos 8 hasta la 11, obtenemos que el término en la posición 1 es caballero, y vemos que no es la palabra buscada, por lo que ahora nos quedamos con las posiciones superiores.

Caemos en la posición 2 del índice, y rearmando el término con los mismos pasos vemos que es “caballo” el cual estábamos buscando.

Leemos desde la posición 10 hasta la 14 del listado de documentos.

Obtenemos que “caballo” aparece en los documentos 1, 2 y 4

Buscamos “caballero”

Accedemos a la mitad de la tabla, ya conocemos que el término es cabaña, accedemos a la posición 1 del índice, y como ya accedimos anteriormente sabemos que el término es caballero, por lo cual encontramos el término buscado.

Leemos desde la posición 6 hasta la 9 del listado de documento.

Obtenemos que “caballero” aparece en los documentos 1 y 4.

Por lo tanto, tenemos que los documentos que devuelve la consulta son los siguientes: 1 y 4.

C)

Los cambios que se deberían hacer es agregar la frecuencia y las posiciones en el listado de los punteros a documentos, esto te permite tener en cuenta las posiciones de los términos en el documento. Por ejemplo, si no se tuviesen las posiciones en el listado, la consulta por frase traería como resultado el documento 1 y 4, cuando el documento 4 es un falso positivo, porque caballo y caballero no son contiguos en ningún momento. Pero, agregando las posiciones no pasaría esto.

Para ilustrar a lo que me refiero dejo un ejemplo de cómo quedaría el listado en el caso de término cabalgar. El formato sería: doc frec posiciones

010 1 00100 010 1 010

Esto se lee como *“En el documento 2 aparece una vez en la posición 4. En el documento 4 aparece una vez en la posición 2”*.