Dante Reinaudo 102848

Information Retrieval 2021-1C 1ra Oportunidad

Dados los siguientes documentos:

- Saca casa
- Aca hay asas
- Casa asa saca
- Aca aca asta

a- Construir un índice invertido que soporte consultas por frase indicando detalladamente los pasos seguidos para su construcción. Almacenar los términos concatenados sin usar front coding y codificar los punteros con códigos gamma.

b- Resolver la consulta por frase “Saca casa” mostrando todos los pasos seguidos para

realizarla. Indicar la cantidad de accesos necesarios para resolverla.

c- Construir una estructura de bigramas, indicando detalladamente cada paso, y resolver la consulta as*a utilizando la estructura creada.

A) Para construir un índice invertido que soporte consultas por frase, debemos almacenar en nuestro listado de documentos concatenados información sobre la frecuencia de cada termino en cada documento, y las posiciones en las que se ubican cada termino en el documento.
Dado que se indica que no utilicemos front coding, para almacenar el léxico utilizaremos directamente el léxico concatenado. Ademas, como etapa de parseo paso todas las palabras a minúsculas al almacenarlas.

Para construir el índice primero debemos recorrer los documentos para obtener todos los términos. Tenemos los siguientes documentos:

- D1: Saca casa
- D2: Aca hay asas
- D3: Casa asa saca
- D4: Aca aca asta

Para cada termino, indicamos en que documentos aparece y su posición relativa en cada documento, comenzando desde la posición numero 1. De esta manera, obtenemos el siguiente listado:

- saca → D1(p1), D3(p3)
- casa → D1(p2), D3(p1)
- aca → D2(p1), D4(p1, p2)
- hay → D2(p2)
- asas → D2(p3)
- asa → D3(p2)
- asta →D4(p3)

Ahora procedemos a ordenar alfabéticamente la lista de términos, ya que estos deben estar ordenados para poder realizar una búsqueda binaria.

- aca → D2(p1), D4(p1, p2)
- asa → D3(p2)
- asas → D2(p3)
- asta →D4(p3)
- casa → D1(p2), D3(p1)
- hay → D2(p2)
- saca → D1(p1), D3(p3)

Ahora procederemos a generar el listado de términos concatenados utilizando léxico concatenado. Para una mejor visualización coloco el índice de la posición donde comienza cada termino debajo de la estructura. Obtenemos los siguiente:

acaasaasasastacasahaysaca

0 3 6 10 14 18 21

Una vez hecho esto procederemos detallar a los documentos a los que pertenece cada termino de modo que se indique su frecuencia y posición:

- Aca → 211 4212
- Asa → 312
- Asas → 213
- Asta → 413
- Casa → 112 311
- Hay → 212
- Saca → 111 313

Luego pasaremos las referencia a los documentos y las posiciones de los términos a distancias:

- Aca → 211 2211
- Asa → 312
- Asas → 213
- Asta → 413
- Casa → 112 211
- Hay → 212
- Saca → 111 213

Ahora codificamos estos valores utilizando código gamma:

- 211 2211 → 01011 01001011
- 312 → 0111010
- 213 → 0101011
- 413 → 001001011
- 112 211→ 11010 01011
- 212 → 0101010
- 111 213→ 111 0101011

Luego generamos la estructura de documentos concatenados, y para una mayor visualización colocaremos índices debajo:

0101101001011 0111010 0101011 001001011 1101001011 0101010 1110101011

0 13 20 27 36 46 53

Para finalizar generamos el índice principal, con dos columnas, una con los punteros al termino en el listado de términos concatenados y otra con los punteros a la estructura de documentos codificados. Obtenemos los siguiente:

Punteros a Terminos	Punteros a documentos
0	0
3	13
6	20
10	27
14	36
18	46
21	53

Entonces nuestro índice invertido este compuesto por las 3 estructuras mostradas en este punto, la tabla con los punteros, el léxico concatenado y los documentos codificados concatenados.

b) Resolver consulta: “Saca Casa”

Para resolver la consulta de la frase “Saca Casa” primero debemos resolver la consulta de cada termino por separado:

Consulta Saca:

1) Debemos realizar una búsqueda binaria, primero leemos la posición 3 del índice, que es la que se encuentra en el medio y obtenemos (10,27). En este primer paso realizamos una lectura del índice. (LI)
2) Ahora leemos la posición siguiente del índice para saber hasta dónde leer en la estructura de términos concatenados, realizamos otra lectura del índice. Obtenemos de la posición 4: (14,36). (LI)
3) Vamos a la estructura de términos concatenados y leemos de la posición 10 hasta la 13. Obtenemos la palabra “asta”. Aquí realizamos una lectura de la estructura de términos. (LT)
4) Como “asta” es menor a “saca”, vamos a la mitad superior del índice. Al dirigirnos a la posición 5 del índice obtenemos: (18,46). En este paso se realiza otra lectura del índice. (LI)
5) Leemos la siguiente posición del índice para saber hasta dónde tomar de en la estructura de términos concatenados, realizamos otra lectura del índice. Obtenemos de la posición 6: (21,53). (LI)
6) Vamos a la estructura de términos concatenados y leemos de la posición 18 hasta la 20. Obtenemos la palabra “hay”. Aquí realizamos una lectura de la estructura de términos. (LT)
7) Como “hay” es menor que “saca”, vamos a la mitad superior del índice. Nos dirigimos a la última posición, la sexta, la cual ya leímos previamente y considero que podría estar en cache. (LICH)
8) Como es el final del índice leemos desde la posición 21 hasta el final de la estructura de términos concatenados. De esta manera recuperamos el término “saca”, el cual es el buscado. (LT)
9) Ahora vamos a la estructura de documentos y leemos desde la posición 53 hasta el final. Obtenemos lo siguiente 1110101011 (LD)
10) Ahora decodificamos el valor obtenido y obtenemos: 111 213
11) Por ultimo pasamos a el valor decodificado de distancias a valor real del documento y obtenemos: 111 313. Esto quiere decir que el termino “saca” se encuentra en la posición 1 del documento 1 y en la posición 3 del documento 3.

Consulta Casa:

1) Nuevamente debemos realizar una búsqueda binaria, primero leemos la posición 3 del índice, que es la que se encuentra en el medio, obtenemos (10,27). Como esto ya lo hicimos en el paso anterior considero que este valor ya está cacheado en memoria, dado que forma parte de la misma consulta. (LICH)
2) Ahora leemos la posición siguiente del índice para saber hasta dónde leer en la estructura de términos concatenados. Obtenemos de la posición 4: (14,36) (LICH)

3) Vamos a la estructura de términos concatenados y leemos de la posición 10 hasta la 13. Obtenemos la palabra “asta”. Aquí realizamos una lectura de la estructura de términos. (LTCH)
4) Como “asta” es menor a “casa”, vamos a la mitad superior del índice. Al dirigirnos a la posición 5 del índice obtenemos: (18,46). (LICH)
5) Leemos la siguiente posición del índice para saber hasta dónde tomar de en la estructura de términos concatenados, realizamos otra lectura del índice. Obtenemos de la posición 6: (21,53). (LICH)
6) Vamos a la estructura de términos concatenados y leemos de la posición 18 hasta la 20. Obtenemos la palabra “hay”. (LTCH)
7) Como “hay” es mayor que “casa”, vamos a la mitad inferior. Nos dirigimos a la posición 4, la cual ya leímos previamente y considero que podría estar en cache. (LICH)
8) Vamos a la siguiente posición del índice para saber hasta dónde leer, la posición 5 que ya leímos previamente, la cual es (18,46). (LICH)
9) Vamos a la estructura de términos concatenados y leemos de la posición 14 hasta la 17. Obtenemos la palabra Casa que es la buscada. (LT)
10) Como ya leímos la posición 5, sabemos que debemos tomar de la estructura de documentos desde la posición 36 a 45. (LICH)
11) De la estructura de documentos, leyendo desde la posición 36 a 45, obtenemos lo siguiente 1101001011. (LD)
12) Ahora decodificamos el valor obtenido y obtenemos: 112 211
13) Por ultimo pasamos a el valor decodificado de distancias a valor real del documento y obtenemos: 112 311. Esto quiere decir que Casa se encuentra en la posición 2 del D1 y la posición 1 del D3.

Una vez realizada las consultas de ambos términos vemos que los documentos candidatos de contener la frase son D1 Y D3. Analizamos las posiciones relativas de los términos en los documentos y vemos que el documento que contiene la frase “saca casa” únicamente es el documento 1.

En total realizamos las siguientes lecturas:

- 4 lecturas del índice + 8 lecturas repetidas
- 3 lecturas a la estructura de términos concatenados + 2 lecturas repetida
- 2 lecturas a la estructura de documentos

C) Construir una estructura de bigramas, indicando detalladamente cada paso, y resolver la consulta as*a utilizando la estructura creada.

Para crear esta estructura primero separo todos los términos en bigramas:

- T1: aca → $a ca a$
- T2: asa → $a sa a$
- T3: asas → $a sa as a$
- T4: asta → $a as st ta a$
- T5: casa → $c ca as sa a $
- T6: hay → $h ha ay y$
- T7: saca → $s ac ca a$

Ahora separo por bigramas e indico en qué términos aparecen:

- $a → aca, asa asas asta
- ca → aca, casa
- a$ → aca, asa, asta, casa, saca
- sa → asa, asas, casa ,saca
- as → asa, asas, asta, casa
- st → asta
- ta → asta
- $c → casa
- ca →casa
- $h → hay
- ha → hay
- ay → hay
- y$ → hay
- $s → saca
- ac → saca

Por la falta de tiempo no llego a armar la tabla, pero paso a describir como debería seguir:

Una vez tenido el listado de bigramas, debería ordenarlo alfabéticamente y asociar cada bigrama a los índices de los términos que los contienen en la estructura de términos concatenados. Como los bigramas son de longitud fija no es necesario crear una estructura concatenada de ellos. Con esto armo la tabla de dos columnas, una con los bigramas y otra con los punteros a los términos que contienen dicho bigrama.

Para resolver la consulta: as*a, debemos buscar los términos que contengan los bigramas $a AND as AND a$. Hacemos una consulta en la tabla de bigramas por cada bigrama y así obtenemos los términos candidatos de cada bigrama. Luego hacemos un AND entre todos los términos candidatos y obtenemos los términos que resuelven la consulta as*a, que en nuestro caso serían asa y asta.