Parcial 10/06 Information Retrieval

Nota de corrección: El punto c) Esta mal. Se corrigió en clase y hay en piazza una diapositiva de la correción

a) Sean los siguientes documentos y su contenido (considero todas las palabras en minúsculas)
D1: saca casa
D2: aca hay asas
D3: casa asa saca
D4: aca aca asta

Recorremos todos los documentos y armamos un archivo auxiliar con los términos y el documento donde se encuentra dicho termino.
Tambien tomamos la posicion de donde fue extraído. En nuestro caso el archivo sería así

- saca->1(1)
- casa->1(2)
- aca->2(1)
- hay->2(2)
- asas->2(3)
- casa->3(1)
- asa->3(2)
- saca->3(3)
- aca->4(1)
- aca->4(2)
- asta->4(3)

Luego ordenamos dichos terminos por término y documento. Teniendo en cuenta la posicion y la frecuencia. Para visualizarlo se muestra de la forma DocFrecuenciaPosiciones. En nuestro caso quedaría de la siguiente forma

- aca->211
- aca->4212
- asa->312
- asas->213
- asta->413
- casa->112
- casa->311
- hay->212
- saca->111
- saca->313

Unificamos los términos

- aca->211-4212
- asa->312
- asas->213
- asta->413
- casa->112,311
- hay->212
- saca->111,313

Codificamos los documentos y frecuencias a distancias

- aca->211,2211
- asa->312

- asas->213
- asta->413
- casa->112,211
- hay->212
- saca->111,213

Con esto podemos armar el *índice*.

Armamos el léxico sin front coding

El puntero de cada término apunta a los términos concatenados en disco

aca	asa	asas	asta	casa	hay	saca
0	3	6	10	14	18	21

Resta armar el índice de documentos. Para ello vamos a codificar los documentos, frecuencias y posiciones de cada término en Gamma. Tener en cuenta que está codificado con distancias por documentos y frecuencias.

También que los bits estan concatenados, el espacio es para visualizar mejor

Ejemplo de codificación en gamma. g(1) = 1, g(2)=010 , g(4) = 00100, etc

010 1 1 010 010 1 1	011 1 010	010 1 011	00100 1 011	1 1 010 010 1 1	010 1 010	1 1 1 010 1011
0	13	20	27	36	46	54

Unificando los índices quedaría de la siguiente manera

punteros terminos	punteros documentos
0	0
3	13
6	20
10	27
14	36
18	46
21	54

aca	asa	asas	asta	casa	hay	saca
0	3	6	10	14	18	21

010 1 1 010 010 1 1	011 1 010	010 1 011	00100 1 011	1 1 010 010 1 1	010 1 010	1 1 1 010 1011
0	13	20	27	36	46	54

b) Resuelvo la consulta "Saca casa", para ello utilizo de ayuda un indice de posición de la tabla

posicion tabla	punteros terminos	punteros documentos
0	0	0
1	3	13
2	6	20
3	10	27
4	14	36
5	18	46
6	21	54

aca	asa	asas	asta	casa	hay	saca
0	3	6	10	14	18	21

010 1 1 010 010 1 1	011 1 010	010 1 011	00100 1 011	1 1 010 010 1 1	010 1 010	1 1 1 010 1011

saca

- Pos 3 (asta) (1 indice, 1 disco) Leemos 4 char: asta. Como saca>asta, busco en la mitad superior
- Pos 5 (hay) (1 indice, 1 disco). Leemos 3 char: hay. Como saca>hay, busco en la mitad superior
- Pos 6 (asta) (1 indice,1 disco). Leemos 4 char: saca.

Saca es lo que buscaba, el puntero a documentos se encuentra en la posicion 54. Leo 10 bits Leo el puntero a documentos en la posición 54 y obtengo 11101010111 en gamma.(1 disco) Que es 111,213 codificado en distancias que es 111,313

casa

- Pos 3 (asta) (1 indice) Ya lo habiamos leído. Como casa>asta, busco en la mitad superior
- Pos 5 (hay) (1 indice). Ya lo habíamos leído también. Como casa<hay, busco en la mitad inferior
- Pos 4 (casa) (1 indice, 1 disco). Leemos 4 chars: casa. casa es lo que buscaba, leo en indice en la posición siguiente para saber el largo del puntero a documentos de casa (1 indice) y tiene largo 10. Leo el puntero a documentos en la posición 36 y obtengo 1101001011 en gamma.(1 disco) Que es 112,211 codificado en distancias que es 112,311

Queda ver en que documentos coinciden ambos terminos y la posición relativa de cada uno en cada documento.
En este caso en el documento 1 coincide pues 111 (saca) y 112(casa) indican que se encuentran en el documento 1 en las posiciones 1 y 2 respectivamente. El resto no cumplen lo solicitado

c) La estructura de bigramas consiste en utilizar una estructura auxiliar que nos permite realizar consultas con comodín de una manera más eficiente

Para cada término de los documentos, los dividiremos en bigramas. Por ejemplo casa sería: "-c ca as sa a-".
Para cada uno de estos bigramas apuntaran a los términos que lo contengan. Por ejemplo:

bigrama	puntero a terminos
-c	4(casa)
as	4(casa),3(asa)..
sa	4(casa),3(asa)..
a-	4(casa),3(asa)..
...	...

Para la búsqueda "as*a" realizamos lo siguiente
Buscamos en el indice que cumpla lo siguiente: -a AND sa AND a-. Es decir buscamos en el índice aquellos que empiecen por a, tengan sa y terminen con a.
Luego realizamos una búsqueda normal como hacíamos con el índice invertido. Es decir, seguimos desde el punto en donde íbamos a buscar a punteros por documento. Reemplazaría al proceso de búsqueda binaria en el índice invertido. Pero solo lo utilizariamos para las búsquedas con comodín.