

7506-2021 Parcialito de IR

Román Vázquez Lareu

TOTAL POINTS

92 / 100

QUESTION 1

1 IR 92 / 100

item a

✓ - **0 pts** ok

item b

✓ - **0 pts** ok

item c

✓ - **8 pts** El índice de bigramas referencia al léxico
cuando debería referir al índice principal.

1 Estos deberían ser posiciones en el índice
invertido principal.

2 Debería haber hecho eso.

3 Si apuntas al índice en vez de a los términos, te
evitas esto.

PARCIAL

D1: Sará casa

D2: Aca hay azar

D3: Casa ará raca

D4: Aca ará rata

a

Extraigo terminos:

Terminos Dic Pos

Sara 1 (1)

Aca 2 (1)

Casa 1 (2)

Aca 4 (1)

Aca 2 (1)

Aca 4 (2)

Aca 2(1), 4(1,2)

hay 2 (2)

ara 3 (2)

ara 3 (2)

azar 2 (3)

ordeno

azar 2 (3)

azar 2 (3)

casa 3 (1)

ordeno

arta 4 (3)

arta 4 (3)

ara 3 (2)

casa 1 (2)

casa 1 (2), 3 (1)

raca 3 (3)

casa 3 (1)

hay 2 (2)

Aca 4 (1)

hay 2 (2)

raca 1 (1), 3 (3)

azar 4 (2)

raca 1 (1)

arta 4 (3)

raca 3 (3)

Aca 2(1), 2(1,1) → 2 1 1 2 2 1 1

~~Doc~~
Doc
Freq
Per

Ara 3(2) → 3 1 2

ara 2(3) → 3 1 3

⇒ aña 4(3) → 4 1 3

distribución
(Doc, per) (ara 1(2), 2(4)) → 1 1 2 2 1 1

Hay 2(2) → 2 1 2

saca 1(1), 2(3) → 1 1 1 2 1 3

Archivo léxico concatenado (sin front coding)

ACA ASA ASASASTA CASAHAYSACA

Pos: 0 3 6 10 14 18 21 24 bytes

Código γ: $\gamma(1) = 1$

$\gamma(2) = 010$

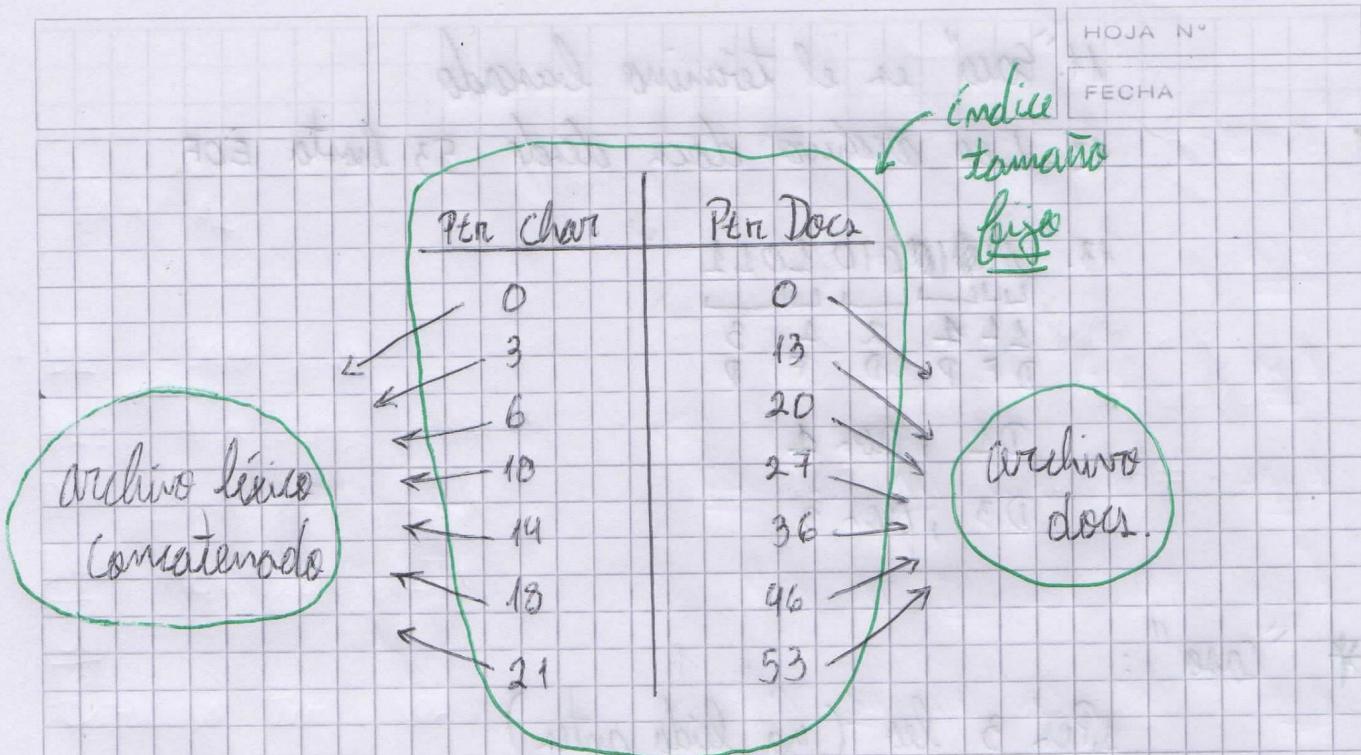
$\gamma(3) = 011$

$\gamma(4) = 00100$

Archivo de Doc:

Pos: 0 13 20 27
01011010010111 0111010 0101011 001001011

Pos: 96 46 53
11010001011 0101010 11010101011 62 bits



El producto final serán estas tres estructuras

b) Q = "Saca cosa"

Primero hago consultas puntuales

* "Saca":

1. Por 3 del índice (1 índice, 1 índice)
2. Leo entrada siguiente para ver hasta donde leo (1 índice)
3. Leo desde [10, 14).
4. "Saca" > "asta" \Rightarrow mitad superior en búsqueda binaria
5. Por 5 del índice (1 índice, 1 índice)
6. Leo entrada siguiente para ver hasta donde leo (1 índice)
7. Leo desde [18, 21)
8. "Saca" > "luya" \Rightarrow mitad sup.
9. Por 6 del índice (1 índice, 1 índice)
10. Leo desde 21 hasta EOF

11. "Sola" es el término buscado

12. Leo archivo docx desde 53 hasta EOF

13. ~~11 10 010 1011~~

~~| |~~
~~1 1 4 2 1 3~~
~~D F P D F P~~

D1 , por 4

D3 , por 3

* "Casa":

1. Por 3 leo (ya leída anterior)

2. Leo entrada siguiente para ver hasta donde leo

3. Leo desde [10, 14]

4. "Casa" > "arta" \Rightarrow mitad superior

5. Por 5 leo (ya leída anterior)

6. Leo entrada siguiente para ver hasta donde leo
(ya leída anterior)

7. Leo desde [18, 21]

8. "Casa" < "ray" \Rightarrow mitad inferior

9. Por 4 (+ índice, 1 disco)

10. Leo entrada siguiente para ver hasta donde voy

11. Leo desde [14, 18)

12. Ese es el término buscado

13. Leo entrada siguiente para ver cuantas veces en el archivo de docx

14. Leo desde [36, 46)

15. 1101001011
 44 | | 44
 11 2 2 11
 D F P D E D

D1, pos 2

D3, pos 1

Hago intersección entre documentos, conservando D1 y D3

Doc 1

"Saca" pos 1 }
 "Casa" pos 2 } Match ✓

Doc 2

"Saca" pos 3 }
 "Casa" pos 1 } No match ✗

⇒ Devuelvo D1.

- Al indexar el término, construyo índice de bigramas, cada entrada del índice es un bigrama y contiene puntero a la lista de términos que lo contienen

~~Secuencia \$S sa ca ca at~~

Aca: \$ A Ac ca at

Ara: \$ A Az ra at

Araa: \$ A Az ra ca z

Asta: \$ A Az st ta at

Casa: \$ C Ca as sa at

hay: \$ h ha ay g

Saca: \$ S sa ca ca at

Los bigramas
tienen largo
fijo.

bigrama] per → podrían ser posiciones en archivo
donde tengo los listos concatenados

\$A	→ [Aca, Ara, Araa, Asta]
Ac	→ [Aca, Saca]
Az	→ [Ara, Araa, Asta]
ay	→ [hay]
at	→ [Aca, Ara, Asta, Casa, Saca]
C	→ [Casa]
Ca	→ [Aca, Casa, Saca]
h	→ [hay]
ha	→ [hay]
S	→ [Saca]
sa	→ [Ara, Araa, Casa]
st	→ [Asto]
ra	→ [Araa]
ta	→ [Asto]
NOTA	g → [hay]

$$Q = as^*a \rightarrow \$a \text{ AND } as \text{ AND } a\$$$

Bigramas de Q

Ahora obtengo la lista de términos de cada bigrama. Supongo esto podría hacerse en búsqueda binaria ya que los ~~bigramas~~ bigramas están ordenados.

$$\$a \rightarrow [Aca, Aza, Ara, Asa]$$

$$as \rightarrow [ara, Aza, Asa]$$

$$a\$ \rightarrow [Aca, Aza, Asa, cara, soca]$$

Hago la intersección de las listas.

$$[ara, Asa]$$

Me fijo que matchean y además se corresponden con la búsqueda.

Una vez obtenidos los términos, puedo hacer la búsqueda puntual de cada uno con el índice invertido del punto ³ (a). Ocupa mucho espacio esta solución.

El resultado final sería

$$D3 \rightarrow Ara$$

$$D4 \rightarrow Asa$$

* Ara:

1. Por 3 del índice invertido
2. Leo hasta siguiente por para ver cuanto leo
3. Leo desde [10, 14)
4. "Ara" < "Anta" \Rightarrow mitad inf
5. Leo por 1
6. Leo siguiente entrada para ver hasta donde leo
7. Leo desde [3, 6)
8. Es el término buscado
9. Leo entrada siguiente para ver hasta donde leo de los dos [13, 20)
10.
$$\begin{array}{r} 0111010 \\ \hline 3 \quad 4 \quad 1 \quad 2 \end{array}$$
D 3 , por 2

* Anta :

1. Por 3 (ya leído)
2. Leo siguiente para ver hasta donde leo
3. Leo desde [10, 14) (ya leído)
4. Es el término
5. Leo entrada siguiente para ver cuanto leo en Dic
6. Leo desde 27 a 36

NOTA

7. 001001011
4 2 3

D₄, para 3

~~desarrollar~~ D₃ y D₄ y las
respectivas posiciones

1 IR 92 / 100

item a

✓ - 0 pts ok

item b

✓ - 0 pts ok

item c

✓ - 8 pts El índice de bigramas referencia al léxico cuando debería referir al índice principal.

1 Estos deberían ser posiciones en el índice invertido principal.

2 Debería haber hecho eso.

3 Si apuntas al índice en vez de a los términos, te evitas esto.