CS 533 Assignment #6

Abdullah Bulbul

Q1)

a) Cover coefficient based

D matrix = $\begin{bmatrix}1 & 1 & 0 & 0\\ 1 & 1 & 0 & 0\\ 0 & 0 & 1 & 1\\ 0 & 0 & 1 & 1\end{bmatrix}$, all α values are ½

S, S^{-1} = $\begin{bmatrix}0.5 & 0.5 & 0 & 0\\ 0.5 & 0.5 & 0 & 0\\ 0 & 0 & 0.5 & 0.5\\ 0 & 0 & 0.5 & 0.5\end{bmatrix}$ $\Rightarrow$ C = $\begin{bmatrix}0.5 & 0.5 & 0 & 0\\ 0.5 & 0.5 & 0 & 0\\ 0 & 0 & 0.5 & 0.5\\ 0 & 0 & 0.5 & 0.5\end{bmatrix}$ $\Rightarrow n_c = 2$

without term1 or term2

D matrix = $\begin{bmatrix}1 & 0 & 0\\ 1 & 0 & 0\\ 0 & 1 & 1\\ 0 & 1 & 1\end{bmatrix}$, all α values are ½

S = $\begin{bmatrix}1 & 0 & 0\\ 1 & 0 & 0\\ 0 & 0.5 & 0.5\\ 0 & 0.5 & 0.5\end{bmatrix}$, S^{-1} = $\begin{bmatrix}1 & 1 & 0 & 0\\ 0 & 0 & 0.5 & 0.5\\ 0 & 0 & 0.5 & 0.5\end{bmatrix}$ $\Rightarrow$C = $\begin{bmatrix}1 & 1 & 0 & 0\\ 1 & 1 & 0 & 0\\ 0 & 0 & 0.5 & 0.5\\ 0 & 0 & 0.5 & 0.5\end{bmatrix}$ $\Rightarrow n_{c1} = 3$

term3 and term4 give a similar situation.
For terms 2, 3, and 4 number of clusters are 3 and TDV is -1 (=2-3).

So, TDV array = [-1,-1,-1,-1]

b) Similarity based

S = $\begin{bmatrix}1 & 1 & 0 & 0\\ \times & 1 & 0 & 0\\ \times & \times & 1 & 1\\ \times & \times & \times & 1\end{bmatrix}$ $\Rightarrow$ Q = 1/3

without t1 (also without t2, t3, or t4) the similarity matrix and Q_1 are the same.

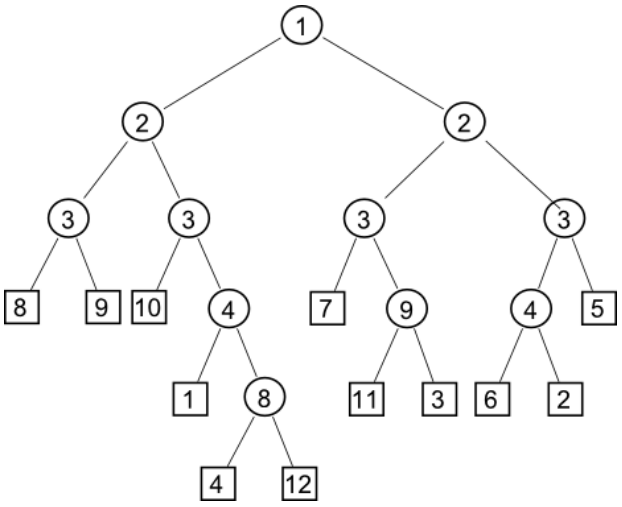
So, TDV array = [0,0,0,0]

Q3)

a) Sistrings:

```
 1: 011011100010111000
 2: 11011100010111000
 3: 1011100010111000
 4: 011100010111000
 5: 11100010111000
 6: 1100010111000
 7: 100010111000
 8: 00010111000
 9: 0010111000
10: 010111000
11: 10111000
12: 0111000
```

PAT Tree




```
PAT array = [8,9,10,1,4,12,7,11,3,6,2,5]
```

b) We can get the indices of A and B using the PAT tree. Then the pairs of indices corresponding to A and B positions which has a smaller difference than d

d = 20 + length of A, for A<20>B

d = 20 + length of B, for B<20>A

should be retrieved.

Q5)

10240 documents

signature size = 192 bits

a) Sequential signature file:

Number of pages to keep signature files $= \frac{10240 \times 192}{0.25 \times 8 \times 1024} = 960$

All pages should be accessed, so the answer is 960

b) Bit-sliced signature file:

For a single bit of signature 10240 bits are needed. (one for each document)

A single bit can be kept in $\frac{10240}{0.25 \times 8 \times 1024} = 5$ pages.

We need 5 page accesses for each bit of query. So number of page accesses is 5 * 5 = 25.

Q7)

a)

```
S1: 0110  1100
S2: 1010  0011
S3: 0011  1100
S4: 0000  1111
S5: 1011  0100
S6: 0100  1011
```

00	01	10	11
S3,S4	S1,S6	S2,S5	-

b)

	activated parts	PAR	SAR
Q1: 0001 1110	4, all	4/4	6/6
Q2: 1001 1100	2, (10,11)	2/4	2/6
Q3: 0011 0011	4, all	4/4	6/6
Q4: 1100 1100	1, (11)	1/4	0/6

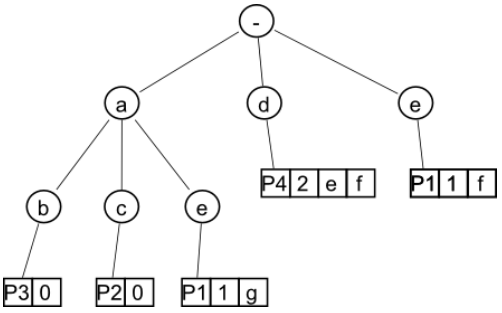
Q9)

a)

P1	a, e, g
P2	a, c
P3	a, b, d
P4	d, e, f
P5	e, f

b) a, e, f, g, b (a, e, g, f are already P1 and P5, if the other term was c P2 would have been satisfied too, so it is b)

c)



d) Only the indicated subtree (on the right) should be searched