Name:- Sumit Kumar Yadav
Roll No.:- 18CS30042

Page No.
Date

Soln:- 1:- 776, 801, 1101, 3125 13

(i) variable byte Code:-

776

⇒ 00000110
10001000

1 10001000

now for gaps: 801 − 776 = 25

⇒ 10001001

for gap: 1101 − 801 = 300

⇒ 00000010
10101100

10101100

for gap 312513 − 1101 = 311412

00010011
00000000
11110100

00010011 00000000 11110100

(ii) gamma codes:-

776, 25, 300, 311412

for 776: in binary 1100001000

∴ γ code: 1111111110 100001000

for 25: in binary 11001,

γ code: 11110 1001

for 300: in binary 100101100.

γ code: 111111110 000101100

for 311412: binary: 1001100000000111010100.

γ code:- 11111111111111111110, 00110000000011
10100.

(b) Gamma-coded gaps :

0|1111000|11|01|1111|01011|11010|1110111

break like this

~~1111000~~ ⇒ 4

1111|00011 ⇒ 100 11 ⇒ 19

1|01 ⇒ 11 ⇒ 3

11111|0 1|00111 ⇒ 1 10111 ⇒ 55

110|10 ⇒ 110 ⇒ 6

1110|111 ⇒ 1111 ⇒ 15

∴ docID = 0 .

now, sequence of Gaps 3, 55, 6, 15

& first entry will be 19.

& sequence of postings:

⇒ 19, 19+3, 19+3+55, 19+3+55+6,

19+3+55+6+15

⇒ 19, 22, 77, 83, 98 .

Soln:- 3:- no. of terms = 250000

no. of tokens = 40,000,000

Bytes per token = 8

Bytes per term id/doc id = 4

• we have to consider 20 blocks.

(i) reading collection of data from disk

= no. of token × ave. byte per token × transfer time + seek time

$= 4\times10^{7} \times 8 \times 2\times10^{-8} + 5\times10^{-3}$

$= 32\times2\times10^{-1} + 5\times10^{-3}$

$= 6.4 + 0.005$ sec

≈ 6.4 sec

(ii) Total sorting time

= no. of blocks × (N log N) * time for low level

where, N = no. of record in each block

$= \frac{\text{no. of token}}{20}$

$= \frac{4\times10^{7}}{20} = 2\times10^{6}$

∴ Total sorting time

$= 20 \times \frac{2\times10^{6}}{20} \log\left(\frac{2\times10^{6}}{20}\right) \times 10^{-8}$

$= 4\times10^{-1} \times \log(2\times10^{6})$

$= 4\times10^{-1} \times 21$

$= 8.4$ sec.

(iii) Total writing time
= no. of block × time for writing block

$= \left(\frac{25\times10^4}{20} + \frac{4\times10^7}{20}\right) \times 4 \times 2\times10^{-8}$

neglect (very small)

$\approx \frac{4\times10^7}{20} \times 4 \times 2\times10^{-8}$

$= 16\times10^{-2}$

∴ total writing time
$= 20 \times 16\times10^{-2}$
$= 320\times10^{-2} = 3.2$ sec.

(iv) total disk transfer time = Total disk seek time × sorted pass time + time writing.

$= 20 \times 6.4 \times 5\times10^{-3} \times 2$

$= \left(\underbrace{25\times10^4}_{\text{neglect}} + \frac{4\times10^7}{2}\right) \times 4 \times 2\times10^{-8}$

∴ total disk transfer time
$= 1.28 + 32\times10^{-1} \times 2$
$= 1.28 + 3.2 \times 2 =$ ~~9.48 sec~~
$= 7.68$ sec.

(v) time for ~~actual merging~~ $= 7.68 + (10^{-8} \times 25\times10^4)$
$= 7.68 + 25\times10^{-4}$
≈ 7.68 sec.

excluding disk transfer time,
operation time for merging $= 25\times10^4 \times 10^{-8}$
$= 25\times10^{-4}$
$= 25\times10^{-4}$ sec.

Soln 4:-

	doc ID	J1	J2	J1 ∧ J2	J1 ∨ J2
(b)	1	0	0	0	0
	2	0	0	0	0
	3	1	1	1	1
	4	1	1	1	1
	5	1	0	0	1
	6	1	0	0	1
	7	1	0	0	1
	8	1	0	0	1
	9	0	1	0	1
	10	0	1	0	1
	11	0	1	0	1
	12	0	1	0	1
	13	0	1	0	1
	14	1	0	0	1
	15	0	0	0	0

(b) precision $= \frac{2}{6} = \frac{1}{3}$

recall $= \frac{2}{2} = 1$

(c) precision $= \frac{6}{6} = 1$

recall $= \frac{6}{6+6} = \frac{1}{2}$

Page No.
Date

Soln:- 2:-

(a) Adv:- lookup time is faster than tree
i.e: it takes O(1) while tree take O(log n)

Disad:-
In hashing there is no easy way to handle the wild-card queries / prefix search

(**) (c) for searching red*
will use $red* as permuterm wildcard index

(b) There are O(n) extra term will be there by considering each term
ex:- abc ⇒ abc$
⇒ abc$, bc$a, c$ab, $abc

(d) for trigram index:
red*, $red*
~~red~~: ($re ∧ red)

(e) bar & barbaric same
↓
ba ar
ba ar rb ba ar ri ic
∴ Jaccard = 2/5