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

Sol^n 1:-

(a)

(i)

	Doc1	Doc2	Doc3
covid	1	0	1
cases	1	0	1
in	1	1	1
india	1	1	0
are	1	0	0
rising	1	0	0
everyday	1	0	0
a	0	1	0
new	0	1	0
approach	0	1	0
for	0	1	0
vaccination	0	1	0
has	0	1	0
been	0	1	0
developed	0	1	0
may	0	0	1
still	0	0	1
surge	0	0	1
december	0	0	1

(ii) covid → 1 → 3

cases → 1 → 3

in → 1 → 2 → 3

india → 1 → 2

are → 1

rising → 1

everyday → 1

a → 2

new → 2

approach → 2

for → 2

vaccination → 2

has → 2

been → 2

developed → 2

may → 3

still → 3

seuge → 3

december → 3

(b) covid OR vaccine : total size = 326512
+ 233312
= 560124

India OR lockdown : total size = 400530
+ 461658
= 562188

delta OR variant : total size = 107913
+ 87009
= 194922

∴ order of query processing will be in sorted order of the size of the OR operations.

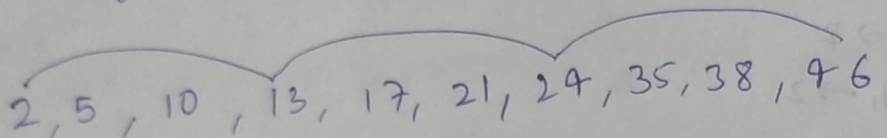
∴ ~~first taken the size of~~

first processing : (delta OR variant) and (covid OR vaccine)

then its result and (india or lockdown).

Solⁿ:- 4(a)

(i) using skip pointers in posting list 1.



4, 10, 18, 25, 35

sequence of comparisons:-

(1) 2&4 (2) 13 & 4 (3) 5 & 4 (4) 5& 10

(5) $\underbrace{10\&10}_{\text{match}}$ (6) 13 & 18 (7) 24 & 18 (8) 17 & 18

(9) 21&18 (10) 21&25 (11) 24&25 (12) 46 & 25

(13) 35 & 25 (14) $\underbrace{35 \& 35}_{\text{match}}$ → now second list finished we stop here

match:- 10, 35

(ii) without the use of skip pointers :-

Sequence of comparisons:-

(1) 2 & 4 (2) 5 & 4 (3) 5 & 10 (4) $\underbrace{10 \& 10}_{match}$

(5) 13 & 18 (6) 17 & 18 (7) 21 & 18 (8) 21 & 25

(9) 24 & 25 (10) 35 & 25 (11) $\underbrace{35 \& 35}_{match}$ → now second list finished we stop here.

match :- 10, 35

(b) words1 /k word2 ⇒ word1 within k words of word2.

q: information /2 retrieval.

So let's consider information positional index

information :-
14 : <36, 174, 252, 651>
23 : <12, 22, 102, 432>
35 : <17>

& retrieval:- ~~2: <6, 78, 1~~

we only have to consider the matched document ie document in which both words present.

∴ doc. no., 14, 23

consider
14: <36, 174, 252, 651> : information
14: <9, 69, 149> : retrieval
no match found for value in range of 2.

now for 23
23: <12, 22, 102, 432> : information
23: <17, 89, 404> : retrieval

(b) information /2 retrieval :-

consider only those document in which both word present.

ie; 2, 23 & 78

now for 2 :-

2 : < 3, 37, 76 >   : information

2 : < 6, 78, 197 >   : retrieval

∴ match found in doc 2 : as (76) /2 (78)

now for 23 :-

23 : < 10, 88, 723 >   : information

23 : < 17, 89, 404 >   : retrieval

∴ match found in doc 23 : as (88) /2 (89)

now for 78 :-

78 : < 15, 25, 195 >   : information

78 : < 10, 23, 198 >   : retrieval

∴ match found in doc 78 as (25) /2 (23)

∴ 3 documents satisfy the given query

ie; document 2, 23 and 78.

Solnː 3:-

Euclidean normalized document vectors:-

(a)

Term	Doc1	Doc2	Doc3
House	0.93	0.1	0.422
flat	0.119	0.615	0
loan	0	0.78	0.80
Discount	0.344	0	0.42

using formula:- $\dfrac{\text{value of tf-idf}}{\sqrt{\Sigma(\ )^2}}$ ↳all terms in particular column.

65.49

50.43

eg:- $\dfrac{15}{\sqrt{(15)^2+(5.2)^2+(40.5)^2}}$

(b) a q: " house loan"

(a) q, Doc1 = 0.93 + 0 = 0.93

q, doc2 = 0.1 + 0.78 = 0.88

q, doc3 = 0.422 + 0.8 = 1.222

∴ rank :- doc2, doc1, doc3

Solⁿ: 2:-

(a) q: "best car insurance"

ntc.ntc.

$\underset{(t_f)}{\downarrow} \quad \underset{\log\left(\frac{N}{df}\right)}{\downarrow}$

df-idf score:-

car : $\log\left(\frac{806791}{18165}\right) = 1.647$

insurance: $\log\left(\frac{806791}{19241}\right) = 1.622$

best : $\log\left(\frac{806791}{25235}\right) = 1.504$

term	doc 1	doc 2	doc 3
car	44.469	6.588	39.528
insurance	0	53.526	47.038
best	21.056	0	25.568

(b) Yes has affect the result like consider example any term occur in any document 10 times that does not mean it is more important than it occur in any document 1 times.

so for this we take log terms