Sahil Jindel 18CS10048

Information Retrieval

Test -1

Ans (a) (i)

Word	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

~~Ans (B)~~

(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
shell	→	3
surge	→	3
December	→	3

Ans (b) ~~for covid OR~~

len (covid OR vaccine) = 326812 + 233312
= 560124

len (India OR lockdown) = 400530 + 161658
= 562188

len (delta or variant) = 107913 + 87009
= 194922

Thus we should first process
(delta OR variant) AND (covid OR vaccine)
and then the result with AND (India OR lockdown)

∴ ((delta OR variant) AND (covid OR vaccine)) AND (India OR lockdown)

Ans2(a) N = 806791

term	df	idf [$\log(N/df)$]
car	18165	1.647
insurance	19241	1.622
best	25235	1.504

The matrix for tf-idf score using ntc, ntc [that is $tf = tf_{t,d}$] is:

	car	insurance	best
d1	27*1.647 = 44.469	0*1.622 = 0	14*1.504 = 21.056
d2	4*1.647 = 6.588	33*1.622 = 53.526	0*1.504 = 0
d3	24*1.647 = 39.528	29*1.622 = 47.038	17*1.504 = 25.568
q	1.647	1.622	1.504

Cosine similarity:

$$(d1, q) = \frac{44.469 * 1.647 + 21.056 * 1.504}{\sqrt{44.469^2 + 21.056^2} * \sqrt{1.647^2 + 1.504^2 + 1.622^2}}$$

$$= 0.7731$$

$$(d2, q) = \frac{6.588 * 1.647 + 53.526 * 1.622}{\sqrt{6.588^2 + 53.526^2} * \sqrt{1.647^2 + 1.504^2 + 1.662^2}}$$

$$= 0.656$$

$$(d3, q) = \frac{39.528 * 1.647 + 47.038 * 1.622 + 25.568 * 1.504}{\sqrt{39.528^2 + 47.038^2 + 25.568^2} * \sqrt{1.647^2 + 1.504^2 + 1.662^2}}$$

$$= 0.979$$

Hence Ranking is

d3
d1
d2

Ans 2(b) The IDF is $\log_b\left(\frac{N}{df}\right)$.

Hence if we increase b from 10, IDf will decrease and vice-versa.
This change will make the new score S_N wrt old score S_o as:

$$S_N = S_o * \log_b(10)$$

* The base of the logarithm won't affect the relative scores of two documents.

A3 (a) Normalised Euclidean vectors:

	Doc 1	Doc 2	Doc 3
House	40.5/43.5 = 0.93	6.6/65.5 = 0.1	21.3/50.4 = 0.42
Flat	5.2/43.5 = 0.12	40.3/65.5 = 0.61	0
Loan	0	51.2/65.5 = 0.78	40.5/50.4 = 0.8
Discount	15/43.5 = 0.34	0	21.2/50.4 = 0.42
length	$\sqrt{40.5^2+5.2^2+15^2}$ = 43.5	$\sqrt{6.6^2+40.3^2+51.2^2}$ = 65.5	$\sqrt{21.3^2+40.5^2+21.2^2}$ = 50.4

(b) (i) Score (Doc1) = 0.93
Score (Doc2) = 0.88
Score (Doc3) = 1.22

Here Ranking = Doc 3
Doc 1
Doc 2

(ii) S(D1) = 0.93 * 1.65 + 0 * 1.62
= 1.53
S(D2) = 1.42
S(D3) = 1.989

Ranking = Doc 3
Doc 1
Doc 2

Ans 4 (a)

L1 : 2, 5, 10, 13, 17, 21, 24, 35, 38, 46

L2 : 4, 10, 18, 25, 35

(i)

1. 2, 4
2. ~~5~~ 13, 4
3. 5, 4
4. 5, 10
→ 5. 10, 10
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. 35, 35
15. 38, 35
~~16~~ 16. 46, 35

end

(ii)

1. 2, 4
2. 5, 4
3. 5, 10
→ 4. 10, 10
5. 13, 18
6. 17, 18
7. 21, 18
8. 21, 25
9. 24, 25
10 35, 25
→ 11 35, 35
12. 38, 35
13, 46, 35

Ans 6 · 2 (at 76, 78)

78 (at 25, 23)