Problem : (May 2019 QP).

Module - 6 :- Text Mining.

Terms to know:

① Term Frequency TF(d,t) measures the association of a term 't' with respect to the given document 'd'.

$$TF(d,t) = \begin{cases} 0 & ; \text{if } freq(d,t) = 0 \\ 1 + \log(1 + \log(freq(d,t))) & ; \text{otherwise} \end{cases}$$

② Inverse Document frequency IDF (t) → represents the scaling factor or the importance of a term 't'.

$$IDF(t) = \log \frac{1 + |d|}{|d_t|}$$

Question:

* Term frequency matrix given in the table shows the frequency of terms document.
- Calculate the TF-IDF value for the term T4 in document 3.

Document/term	T1	T2	T3	T4	T5	T6
D1	5	9	4	0	5	6
D2	0	8	5	3	10	8
D3	3	5	6	6	5	0
D4	4	6	7	8	4	4

We have to calculate TF-IDF value for the term T_4 in D_3.

$$TF(d_3, t_4) = 1 + \log(1 + \log(6))$$

$$= 1.249$$

$$IDF(t_4) = \log \frac{(1+4)}{3}$$

$$= 0.220$$

$$TF\text{-}IDF = 1.249 \times 0.220$$

$$= 0.2760$$

Table 10.5 A term frequency matrix showing the frequency of terms per document.

document/term	t_1	t_2	t_3	t_4	t_5	t_6	t_7
d_1	0	4	10	8	0	5	0
d_2	5	19	7	16	0	0	32
d_3	15	0	0	4	9	0	17
d_4	22	3	12	0	5	15	0
d_5	0	7	0	9	2	4	12

Calculate TF-IDF value for the term t_6 in d_4.