Semester : VI Subject : DAV Academic Year: 2022-2024

TF - IDF → Term Frequency and Inverse Document Frequency :-

Term Frequency - Inverse Document Frequency (TF-IDF) is a widely used statistical method in natural language processing and information retrieval. It measures how important a term is within a document relative to a collection of documents (i.e. relative to corpus).

The formula used to calculate Term frequency :

$$\text{Term Frequency} = \frac{\text{No. of repetation of words in a sentence}}{\text{No. of words in a sentence.}}$$

$$\left.\begin{array}{l}\text{IDF (Inverse Document Frequency)}\end{array}\right\} = \log\left(\frac{\text{No. of sentences}}{\text{No. of sentences containing words}}\right)$$

TF-IDF → TF * IDF

Example :

Consider the given 3 sentences. Calculate TF-IDF for the given sentence.

Sentence 1 → Good Boy.
Sentence 2 → good girl
Sentence 3 → boy gir good.

Semester : VI Subject : DAV Academic Year: 2023-2024

Solution:

TF Table.

	Sen 1	Sen2	Sen3.
good	1/2	1/2	1/3
boy	1/2	0	1/3
girl.	0	1/2	1/2.

→ 1 good word in 3 words. In the given sentence = 1/3.

Apply the TF formula to generate the TF Table. The column consists of 3 sentences. The row consists of unique words ~~sentences~~ of all three sentences.

IDF Table.

Words	IDF
good	log(3/3)
boy	log(3/2)
girl	log(3/2)

⟹ This table is generated by applying IDF formula. There are 3 good word in all the three sentences so we get log(3/3).

TF-IDF ⟹ TF * IDF.

	good	boy	girl.
Sen1	0	1/2 log(3/2) = 0.088	0
Sen2	0	0	1/3 × log(3/2) = 0.088
Sen 3	0	1/3 × log(3/2) = 0.088	1/3 × log(3/2) = 0.088.

⟹ This is the calculated TF-IDF value.