TFIDF :

It is combination of two things.

① Term frequency and ② Inverse document frequency

① Term frequency :=

$$\frac{\text{freq. of term 't' in a documents}}{\text{Total no. of words in that documents.}}$$

"Data science is a combination of ml and dl"

↓

data science combination machine learning deep learning.

Teacher's Signature:.....

Term freq. of "learning" = $\frac{2}{7}$

$\Rightarrow$ ~~28.42%~~ 0.2842

T.f (Data) = $\frac{1}{7}$ = 0.1428

Document frequency $\Rightarrow$

$\frac{\text{No. of documents containing 't'}}{\text{total no. of documents.}}$

(1) NLP needs Deep Learning and machine learning

(2) Data science contains lots of things like NLP

document freq of (NLP) = $\frac{2}{2}$ = 1

D.f of "learning" = $\frac{1}{2}$ = 0.5

Inverse document frequency:

IDF $\Rightarrow$ $\log\left(\frac{1}{\text{Document frequency}}\right)$

$\Rightarrow$ $\log\left(\frac{\text{Total no. of Document}}{\text{no. of Doc. containing 't'}}\right)$

$$IDF(\text{NLP}) = \log\left(\frac{1}{1}\right) = \log(1) = 0$$

$$IDF(\text{learning}) = \log\left(\frac{2}{1}\right)$$

$$TFIDF \Rightarrow TF \times IDF.$$

$$TFIDF(\text{learning}) = \frac{2}{7} \times \log(2)$$

exa!

① learning nlp.

② data science combinatics deep learning machine learning

③ nlp needs machine learning deep learning.

⇓

Unique words:

w. In this case every unique word treated as a feature.

	learning	nlp	data	science	combina°	machine	deep	nee
1)								
2)								
3)								

No. of row will be 3.

Teacher's Signature:.....

1st Documents (TF) (IDF)

TFIDF (learning) = $\frac{1}{2} \log\left(\frac{3}{3}\right)$

$\Rightarrow \frac{1}{2} \times 0$

$= 0$

TFIDF (nlp) $\rightarrow \frac{1}{2} \log\left(\frac{3}{2}\right)$

Doc 2

IFIDF (data) = $\frac{1}{7} \log\left(\frac{3}{1}\right)$

TFIDF (science) = $\frac{1}{7} \log\left(\frac{3}{1}\right)$

TFIDF (combination) = $\frac{1}{7} \log\left(\frac{3}{1}\right)$

TFIDF (machine) = $\left(\frac{1}{7}\right) \log\left(\frac{3}{2}\right)$

TFIDF (deep) = $\frac{1}{7} \log\left(\frac{3}{2}\right)$

~~TFIDF (needs) = $\frac{1}{6} \log\left(\frac{3}{1}\right)$~~

TFIDF (learning) = $\frac{2}{7} \log\left(\frac{3}{3}\right)$

DOC3

TFIDF (NLP) = $\frac{2}{6} \log\left(\frac{3}{2}\right)$

TFIDF (needs) = $\frac{1}{6} \log\left(\frac{3}{1}\right)$

Teacher's Signature:

$$TFIDF(machine) = \frac{1}{6} \log\left(\frac{3}{2}\right)$$

$$TFIDF(deep) = \frac{1}{6} \log\left(\frac{3}{2}\right)$$

$$TFIDF(learning) = \frac{2}{6} \log\left(\frac{3}{3}\right)$$

In this case we are just checking the weightage of the word in their respective documents.

freq of word ↑ ⇒ weightage of word ↓

Drawbacks of TFIDF

1. Curse of dimensition
2. Order is not maintained.
3. It is not understanding actual meaning of the word.

exa!

movie reviews

1. Bad movie
2. Gud movie
3. fabulous movie
4. Awesome movie
5. Average movie

} In this reviews movie is not important where as the word also both

PAGE No.
DATE / /

Similarities betn Count vectorizer and TFIDF

(1) Drawbacks of both are same.
(2) When we initiate their model, the paramm eter of the both are same.

Diff e

Count Vectorizer ⇒ Count the word
TFIDF ⇒ Check the weightage of word.

While model training the diff betn the accuracy of both is around 1-2% but in reality it is very high, so from where we get the good evaluation matrics value we will continue with it.

Sparse Matrix: It is a output matrix of count vectorizer.
In this matrix most of the values are zero.
So we convert this matrix into array.

Data leaking :→
In this case we 1st spit the data after that we perform the operation of preprocessing on both training side and testing side separately.

Teacher's Signature:.....