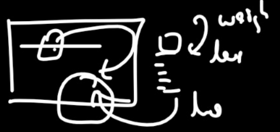
TF - IDF (Term frequency - Inverse Document frequency)

TO evaluate the importance of Word in document or corpus
↓
set

① TF - IDF is a statistical Measure used to evaluate the importance of Word in document relative to a collection of document

② It helps to weigh down common words that occur frequently across documents but are less informative , while giving more weights to rarer words that might be more significant for specific document



$$TF = \frac{\text{No of rep of words in sentence}}{\text{No of words in sentence}}$$

] measures how frequently a word appear in a document

$$IDF = \log_e\left(\frac{\text{No of sentences}}{\text{No of sentences containing the word}}\right)$$

] Measures how important a word in given corpus

IDF ↑ - rare

$$\boxed{\text{TF - IDF = Term frequency} \times \text{Inverse Document frequency}}$$

Example

DOC 1 : " Cat eats fish "

DOC 2 : " dog eats fish "

DOC 3 : " dog chases Cat "

Term frequency

Step-1

Term	DOC-1 (cat eats. fish)	Doc 2 (dog eats fish)	DOC3 (dog chases cat)
cat	$\frac{1}{3}$	0	$\frac{1}{3}$
eats	$\frac{1}{3}$	$\frac{1}{3}$	0
fish	$\frac{1}{3}$	$\frac{1}{3}$	0
dog	0	$\frac{1}{3}$	$\frac{1}{3}$
chases	0	0	$\frac{1}{3}$

$\frac{1}{3}$

Step 2 - Calculate inverse Document frequency

[# of documents containing word)

cat	2
eats	2
fish	2
dog	2
chases	1

$\log\left(\frac{3}{2}\right) = 0.18$

$\log\left(\frac{3}{2}\right) = 0.18$

$\log\left(\frac{3}{2}\right) = 0.18$

$\log\left(\frac{3}{2}\right) = 0.14$

$\log\left(\frac{3}{1}\right) = \boxed{0.48}$

TF-IDF

Doc1

$\text{TF-IDF}(\text{cat}) = \frac{1}{3} \times 0.18 = 0.06$

$\text{eats} = \frac{1}{3} \times 0.18 = 0.06$

$\text{fish} = \frac{1}{3} \times 0.18 = 0.06$

Doc 2

$\text{dog} = \frac{1}{3} \times 0.18 = 0.06$

cats $= \frac{1}{3} \times 0.18 = 0.06$

fish $= \frac{1}{3} \times 0.18 = 0.06$

DOC3

$$\text{TF} \times \underline{\text{IDF}}$$

dog $= \frac{1}{3} \times 0.18 = 0.06$

chew $= \frac{1}{3} \times 0.48 = 0.16$

cat $= \frac{1}{3} \times 0.18 = 0.06$

cat , cats , fish, dog, chew

DOC1 → [0.06, 0.06, 0.06, 0, 0] ←

DOC2 → [0 , 0.06, 0.06, 0.06, 0]

DOC3 → [0.06, 0, 0, 0.06, 0.16]

Advantages

1. Intuitive
2. fixed size
3. Word Importance is getting capture

Disadvantages

1. sparsity
2. growing vocabulary