1. OHE
2. BOW
3. TF-IDF
4. N-GRAM

⇒ frequency based encoding

5. Custom technique

⇒ Vector (text Representation & text encoding)

text → Number

Next → (embedding) → word2vec (Neural Network)

Transformer

- Similirty
  1. Dot Product
  2. Cosine Similary

1. Data ingestion
2. Validate Data
3. Aanalysis
4. Preprocessing ⇒ encoding (embedding) low level
5. Model ← store
6. Evaluate

1. Corpus → Entire text - Paragraph
2. Document - sentance
3. token → word

Vocabulary

Unique collection of word (token) → (corpus) Doc

Paragraph corpus

Improvements in transformer-based deep neural networks enabled an AI boom of generative AI systems in the early 2020s. These include large language model (LLM) chatbots such as ChatGPT, Copilot, Bard, and LLaMA, and text-to-image artificial intelligence art systems such as Stable Diffusion, Midjourney, and DALL-E.[7][8][9] Companies such as OpenAI, Anthropic, Microsoft, Google, and Baidu as well as numerous smaller firms have developed generative AI models.[3][10][11]

OHE

Corpus

= { People watch ineuron ineuron watch ineuron
People write comment- ineuron write comment }

- D_1 People watch ineuron
- D_2 ineuron watch ineuron => 3x6
- D_3 People write comment- =>3x6
- D_4 ineuron write comment- => 3x6

Sunny => 4x6

Document

=> padding

[[000000)
(- - -)
[- - -]]

comment

50000

Sunny

testing krish write

training

(000000)

Vocabulary

[People, watch, ineuron, write, comment-]

6

V=5

OHE

People	watch	ineuron	write	comment-

5

D_1 =

```
[ [1 0 0 0 0]
  [0 1 0 0 0]
  [0 0 1 0 0] ]   (3x5)
```

D_4 =

```
[ [0 0 1 0 0]
  [0 0 0 1 0]
  [0 0 0 0 1] ]   (3x5)
```

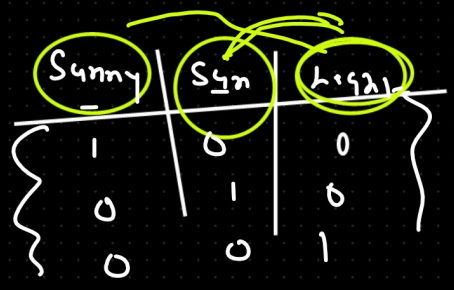
most of the value zero

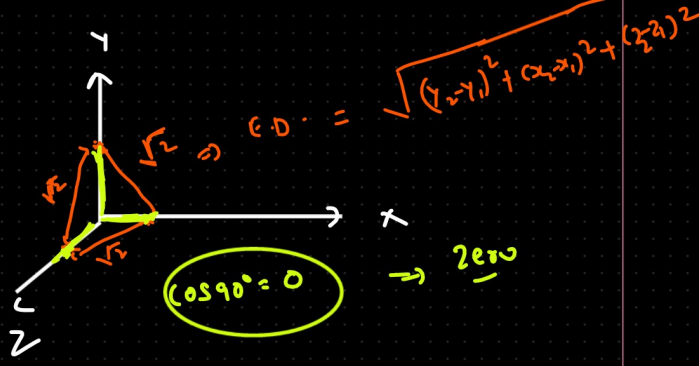
Pros

- Easy to implement

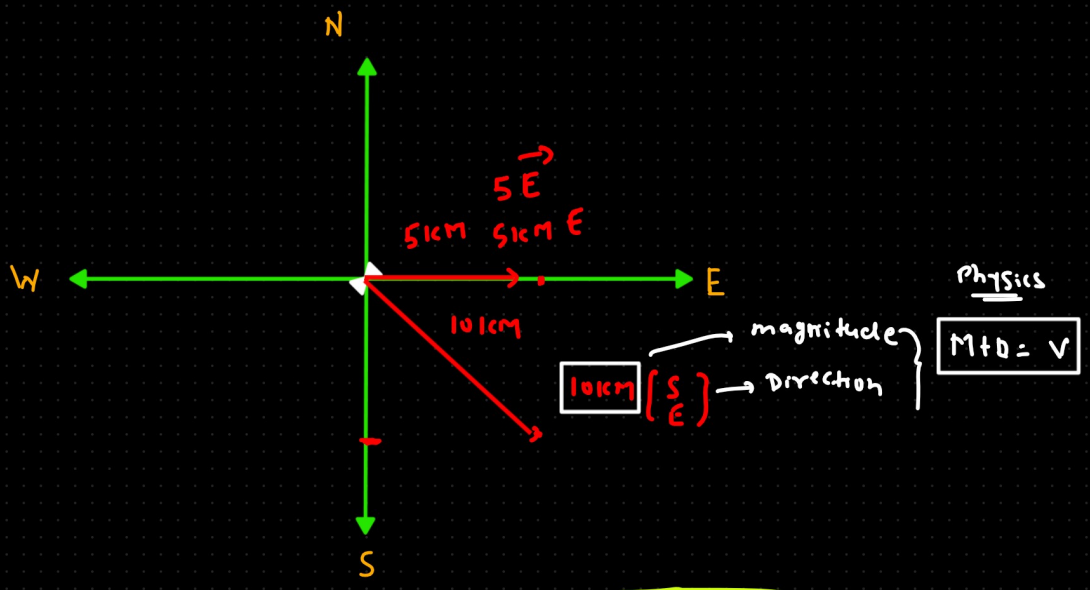
Cons

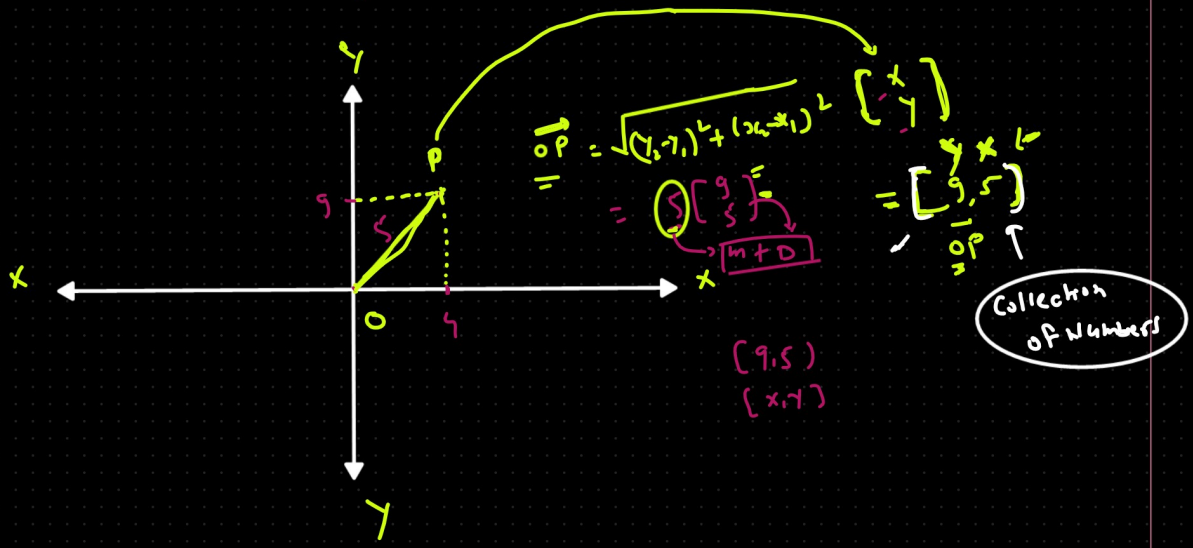
1. Sparsity (Sparse matrix)
2. No fixed size
3. OOV
4. context will not be there

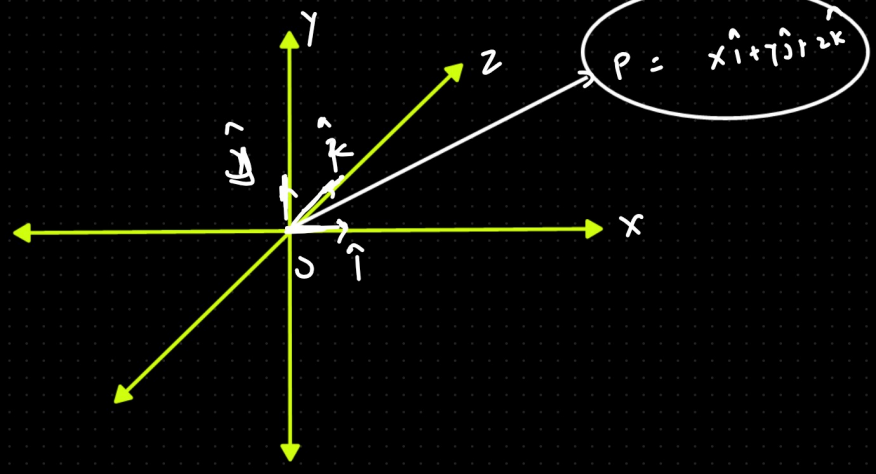

Sunny
Syn
Rainy
1 0 0
0 1 0
0 0 1


Y
X
Z
√2
√2
√2
E.D. = √((x2-x1)^2 + (y2-y1)^2 + (z2-z1)^2)
cos90° = 0
→ zero

Vector


N
W
E
S
5E
5km 5km E
10km
10km [S E]
magnitude
Direction
Physics
M+D = V


Y
X
P
9
5
O
OP = √((x2-x1)^2 + (y2-y1)^2) [x y]
= 5 [9 5]
M+D
= [9 5]
OP
Collection of Numbers
[9,5)
[x,y]


Y
Z
X
O
P = xî + yĵ + zk̂

Bag of word

Watch

Sunny

?

V=5

Sumo

- D_1 People Watch Ineuron
- D_2 Ineuron Watch Ineuron => 3x6
- D_3 People Write Comment => 3x6
- D_4 Ineuron Write Comment => 3x6

training

	People	Watch	Ineuron	Write	Comment
D_1	1	2	1	0	0
D_2	0	1	2	0	0
D_3	1	0	0	1	1
D_4	0	0	1	1	1

Adv

Simple & intuitive

Context

Disadv.

1. Sparsity still is there
2. Order
3. GOV (Dim. issue)

Watch Ineuron

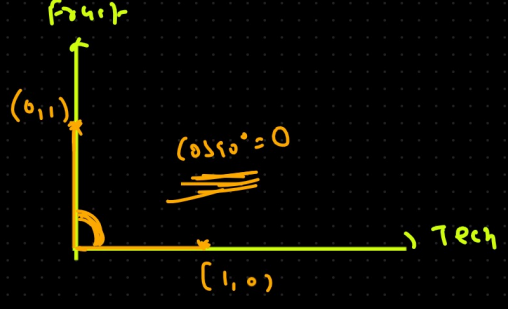
What you are doing?

Apple iPhone

Apple

=>

Tech	Fruit
1	0
0	1



Dot Product

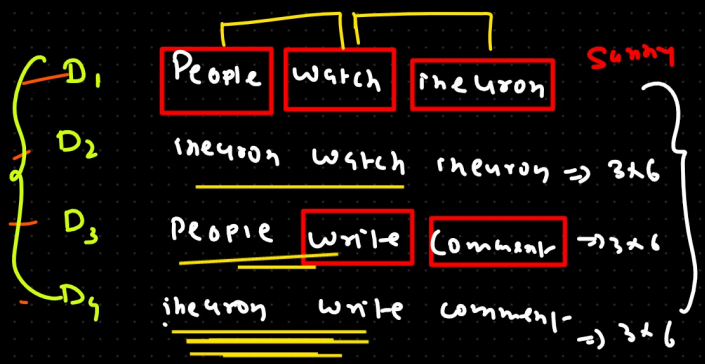
$(x_1, y_1) \cdot (x_2, y_2)$

$(x_1 * x_2) + (y_1 * y_2)$

$(0,1) \cdot (1,0)$

$0*1 + 1*0 = 0$

N-gram



N = 16
N = 5

Single unique word
=>

BOW :) N=1 => Single word — vocab { People watch rneuron write comment }

=> (double unique word)

N=2 = (double word) (bi-gram) => vocab { people watch, watch rneuron, rneuron watch, People write, write comment, rneuron write }

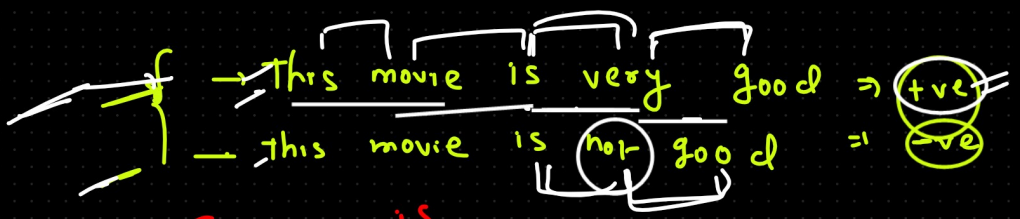
N=3 (tripple word) tri-gram

6 - Unique words

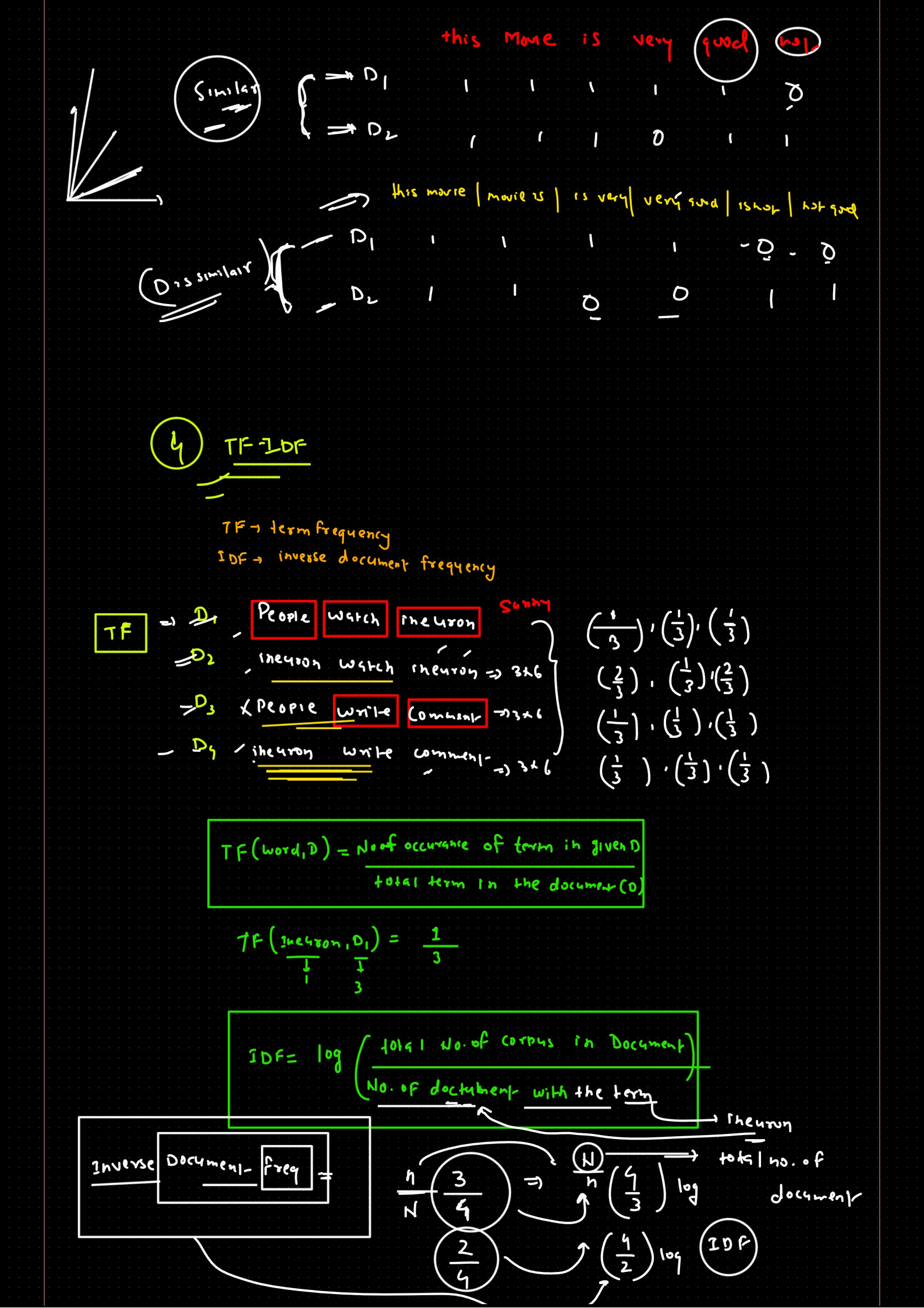
Assignment

	People watch	watch rneuron	rneuron watch	People write	write comment	rneuron write
D_1	1	1	0	0	0	0
D_2	0	1	1	0	0	0
D_3	0	0	0	1	1	0
D_4	0	0	0	0	1	1

Ex:- capturing more context, Reducing the ambiguity



{ this movie is very good not }



4 TF-IDF

TF → term frequency

IDF → inverse document frequency

TF

D1 - People watch ineuron

D2 - ineuron watch ineuron → 3x6

D3 - People write comment → 3x6

D4 - ineuron write comment → 3x6

Sunny

$(\frac{1}{3}), (\frac{1}{3}), (\frac{1}{3})$

$(\frac{2}{3}), (\frac{1}{3}), (\frac{2}{3})$

$(\frac{1}{3}), (\frac{1}{3}), (\frac{1}{3})$

$(\frac{1}{3}), (\frac{1}{3}), (\frac{1}{3})$

$$TF(word, D) = \frac{\text{No of occurance of term in given D}}{\text{total term in the document (D)}}$$

$$TF(\underset{1}{\text{ineuron}}, \underset{3}{D_1}) = \frac{1}{3}$$

$$IDF = \log\left(\frac{\text{total no. of corpus in Document}}{\text{No. of doctument with the term}}\right)$$

ineuron

Inverse Document-freq =

$\frac{n}{N}$ $\frac{3}{4}$ ⇒ $\frac{N}{n}$ $\left(\frac{4}{3}\right)$ log → total no. of document

$\frac{2}{4}$ $\left(\frac{4}{2}\right)$ log → IDF

D_1 People watch theuron

D_2 theuron watch theuron → 3×6

D_3 people write comment → 3×6

D_4 theuron write comment → 3×6

Sunny

$$TF = \frac{\text{freq of given term in doc}}{\text{total term in doc}}$$

$$IDF = \log\left(\frac{\text{total Document}}{\text{Doc No. which contain term}}\right)$$

Alphbatical

(TF×IDF) (TF×IDF) (TF×IDF) (TF×IDF) (TF×IDF)

	People	watch	theuron	write	comment
D_1	$\frac{1}{3}\times\log(\frac{4}{2})$	$\frac{1}{3}\times\log(\frac{4}{2})$	$\frac{1}{3}\times\log(\frac{4}{3})$	0	0
D_2	0	$\frac{1}{3}\times\log(\frac{4}{2})$	$\frac{2}{3}\times\log(\frac{4}{3})$	0	0
D_3	$\frac{1}{3}\times\log(\frac{4}{2})$	0	0	$\frac{1}{3}\times\log(\frac{4}{2})$	$\frac{1}{3}\times\log(\frac{4}{2})$
D_4	0	0	$\frac{1}{3}\times\log(\frac{4}{3})$	$\frac{1}{3}\times\log(\frac{4}{2})$	$\frac{1}{3}\times\log(\frac{4}{2})$

TF = ? → People watch theuron People, People

People → 1/3 theuron → 1/3 People → 3/5

0.33 (0-1) 0.6

D_1 People watch theuron

D_2 People comme theuron

IDF = TF × IDF

$OF = \frac{2}{2} = 1 \rightarrow \frac{1}{1} = 1$

$DF = \frac{1}{2} = 0.5 = \frac{10}{0.5} = 2$ ← IDF

$DF = \frac{2}{2} = 1 = 1$