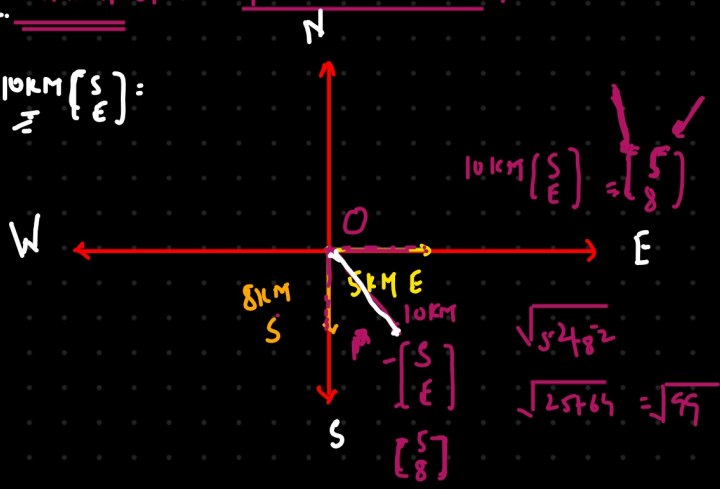
1. Frequency based (encoding)
2. Neural Network based (embedding)

Stemming

Lemmatization

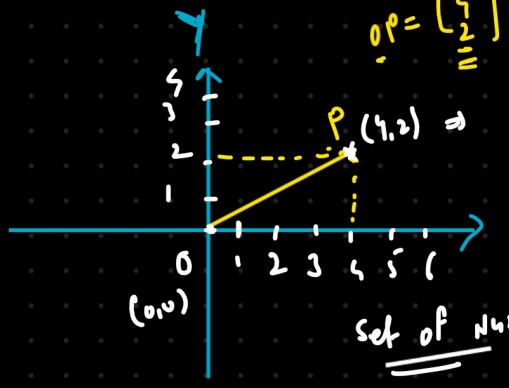
Vector :-

Mathematics ⇒ magnitude + direction



vector :

$\vec{OP} = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$



$OP = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

Euclidean Distance $= \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

$= \sqrt{(1-0)^2 + (2-0)^2}$

$= \sqrt{16+4} = \sqrt{20}$

$OP = \sqrt{20} \begin{bmatrix} x \\ y \end{bmatrix}$

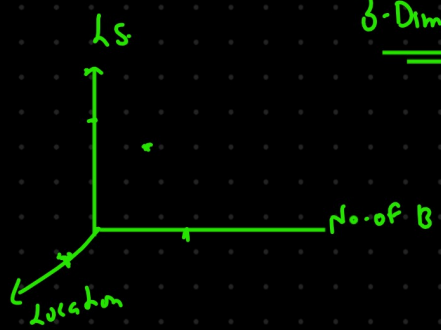
$\vec{OP} = \sqrt{20} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$:

$= \boxed{\vec{OP} = x\hat{i} + y\hat{j} + z\hat{k}}$

House Price

Number of bedroom	housesize	Location	Price
5	1000	B	50L
6	1200	D	
8	1600	M	

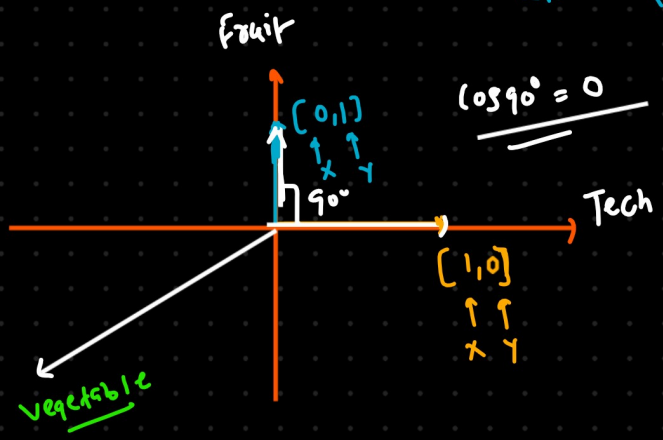
Set of No ⇒ 1-D array



Matrix.

Similarity Search

Apple ⇒ 🍎
apple ⇒ 🍏



3 ⇒ Property

Vector ⇒ 1024 ⇒ No of features
⇒ 384
No. of feature

⇒ Apple = [0, 1] ⇒ (0, 1, 0)
IPHONE = [1, 0] ⇒ [1, 0, 0]
2-D 3D

Apple Iphone

[0, 1] . [1, 0]

Dot Product =

{ Apple, Iphone }

0×1 + 1×0 = 0

Cosine Similarity

$$= \frac{\vec{A} \cdot \vec{B}}{|A| \cdot |B|} = \frac{0}{=} = 0$$

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

} Frequency

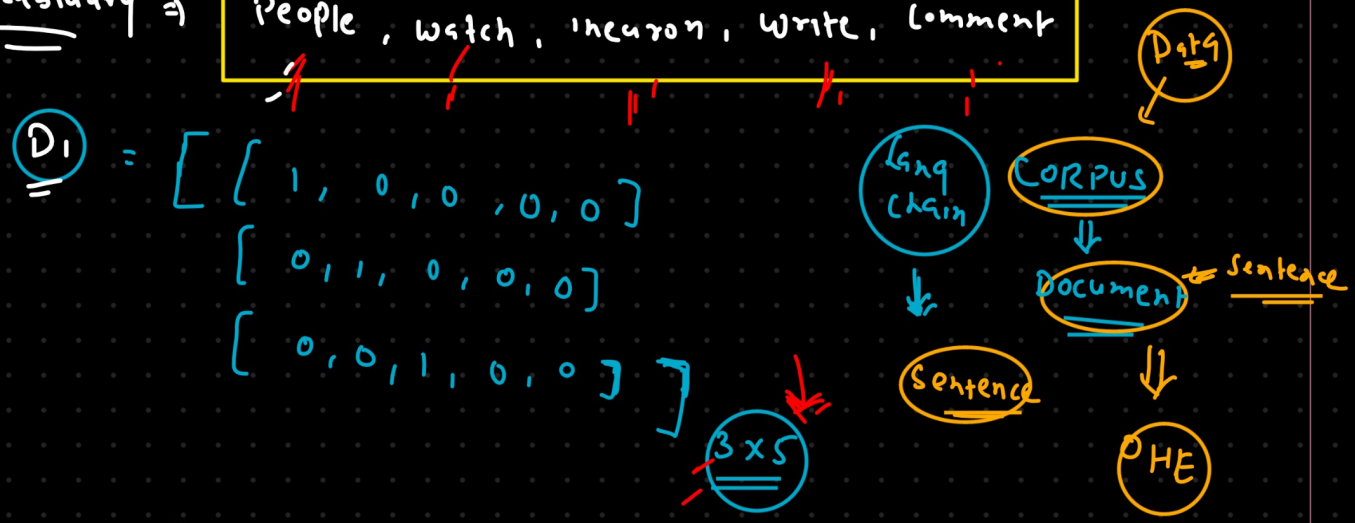
Document → Corpus → OHE

{ People watch ineuron ineuron watch ineuron
People write comment ineuron write comment }

D_1 = People watch ineuron
D_2 = ineuron watch ineuron
D_3 = People write comment
D_4 = ineuron write comment

Many words from corpus
Vocabulary = People, watch, ineuron, write, comment

As many as = Dimension



D_1 = [[1, 0, 0, 0, 0]
[0, 1, 0, 0, 0]
[0, 0, 1, 0, 0]]

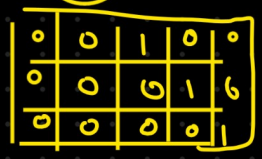
D_2 = [[0, 0, 1, 0, 0]
[0, 1, 0, 0, 0]
[0, 0, 1, 0, 0]]

D_3 = [[1,0,0,0,0] = [0,0,0,1,0] [0,0,0,0,1]]

Word.

old → OHE

D_4 =



D_4 = [[0,0,1,0,0] ← one word
[0,0,0,1,0] ← another word
[0,0,0,0,1]] ← another

= Document → Vocabulary

Disadvantage (2010, 2014)

Sparse (zero)

DL

Syntax Context

② BOW (Bag of words)

BOW ⇒ ① vocabulary
② based on the Document
we calculate Frequency

No → Sunny ⇒ Sunny no. ✗

D_1 → People watch ineuron
D_2 → ineuron watch ineuron
D_3 → People write comment
D_4 → ineuron write comment

	People	watch	Ineuron	write	comment
D_1	1	1	1	0	0
D_2	0	1	2	0	0
D_3	1	0	0	1	1
D_4	0	0	1	1	1

BOW

1. Simple
2. Some sort of context

FEW

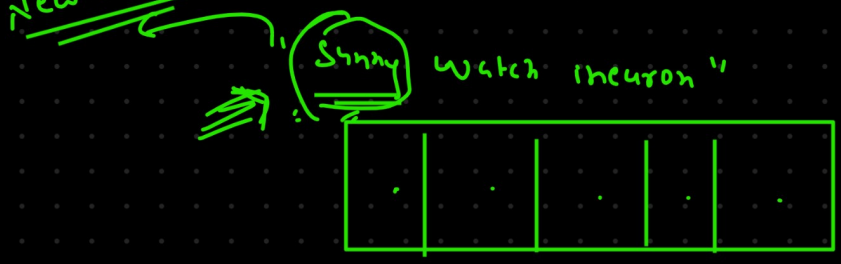
(few no of zero)
Dense

OHE ⇒ Sparse (more zero)

Disadvantage ⇒ 1. if the size of vocab is too long & Parse (zero)
2. OOV (it can not handle)

Doc → Vocablary ⇒ Encode (frequency)

X new context



3. N-GRAMS

N → 2 -
N → 3 -
N → 4 -

N=1 ⇒ unigram
N=3 ⇒ trigram
N=2 ⇒ Bigram ⇒ PAIR of two
" " three

D1 ⇒ People watch neuron
D2 ⇒ neuro watch neuron
D3 ⇒ People write comment
D4 ⇒ neuron write comment

⇒ N=2 ⇒ bi-gram
(Pair of 2 word) Unique
⇒ customized.

N=1 ⇒ { People, watch, neuron, write, comment }

N=2 ⇒ { People watch, watch neuron, neuron watch, People write, write comment, neuron write }

N=3 ⇒ Pair of three words

	People watch	watch neuron	neuron watch	People write	write comment	i between 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

Advantage
1. Capture more context-
2. Reducing the ambiguity

this movie is very good

this move is not good

Context

Unigram N=1 BoW

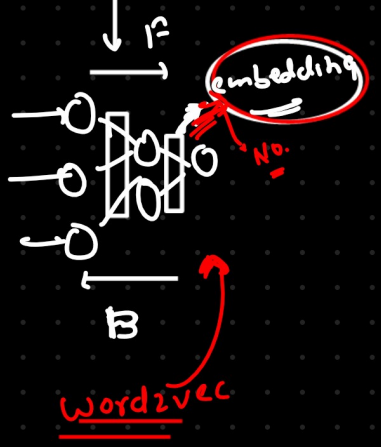
{ This, movie, is, very, good, Not }

N=2

{ This movie, movie is, is very, very good, is not, not good }

N=2

Embedding & Encoding => [ ]



manually (frequency)

BOW (N-grams)

TF-IDF

Word2vec

4 tf-IDF

=> google

TF => term frequency

IDF => Inverse Document- frequency

No of sent = 4

D_1	=>	People	watch	neuron	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$\log(\frac{4}{2}), \log(\frac{4}{2}), \log(\frac{4}{3})$
D_2	=>	neuron	watch	neuron	$\frac{2}{3}, \frac{1}{3}, \frac{2}{3}$	" " "
D_3	=>	People	write	comment	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$\log(\frac{4}{2})$ " "
D_4	=>	neuron	write	comment	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	" " "

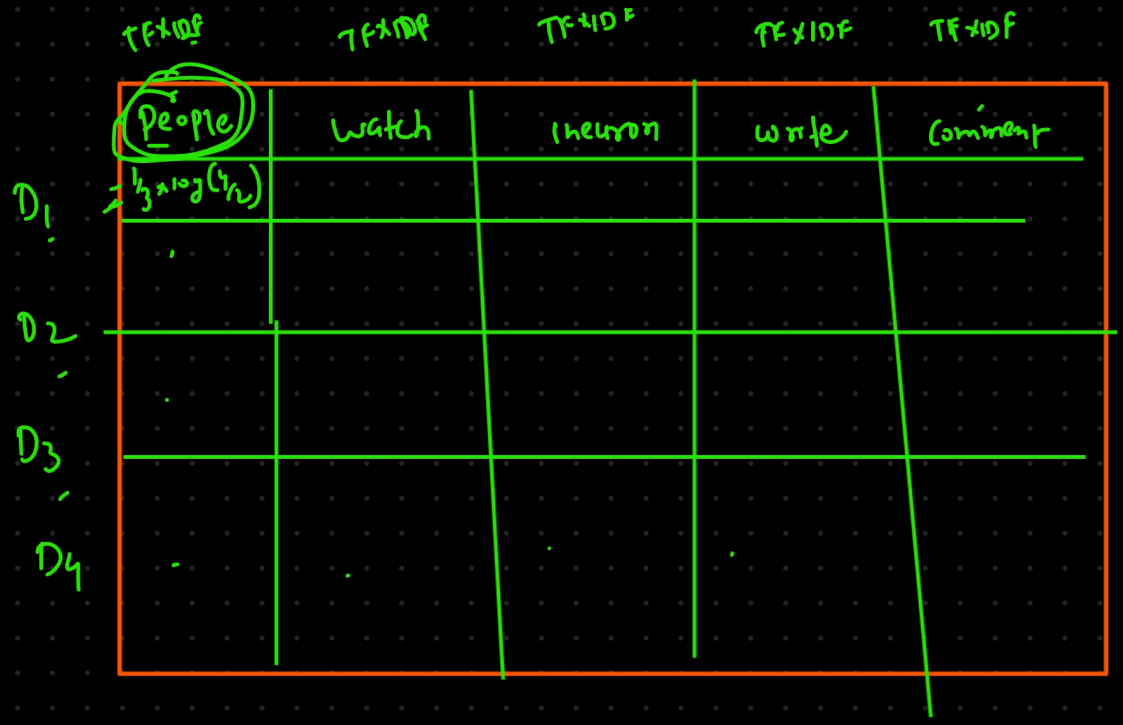
TF => (word i) / (in D_1)

$$TF = \frac{\text{No. of occurance of the words in given D}}{\text{total word in the Document}}$$

$$IDF = \log\left(\frac{\text{Total no. of Document}}{\text{No. of Document which contain given term}}\right)$$

$$\text{neuron} = \log\left(\frac{4}{3}\right)$$

TF x IDF


TF×IDF
TF×IDF
TF×IDF
TF×IDF
TF×IDF
People
Watch
Iheuron
write
Comment
D1
D2
D3
D4
= 1/3 × log(4/2)