- Text cleaning & Text PreProcessing
- Text encoding
- Word embedding =
  - Word2vec
  - BERT embedding
  - Transformer
  - Attention all you need ⇒ Python

Data → Embedding → vector set of No.

1. Vector
2. Vector Similarity
3. Embedding ⇒ Word2vec
4. Python Practical

Pipline

1. Data
2. Process data
3. Model
4. evalute

Mathematics

Text → LLM → text

GPT = 3.5
Gemini = 1.5V

Text → Encoded. → Decoding

Text → Encoding → chatgpt → Decoding → text

Prompt

Word2vec — Neural network

- CBOW
- Skipgram

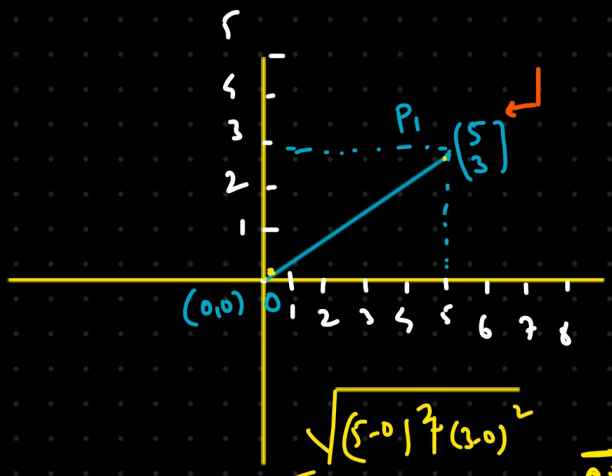
Word2vec ⇒ Word → Vector

1. Semantic meaning
2. Lower Dim
3. Sparsity → Dense

Word2vec

- {Custom model}
- {Pretrain model}
  - 2013 by
  - Google News Article by
  - 3 billion words by
  - 300-features by

[300-Dim] Features



$P_1 = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$ ⇒ vector (set of No.)

2-Dimension

$\sqrt{(5-0)^2+(3-0)^2}$

$= \sqrt{25+9}$

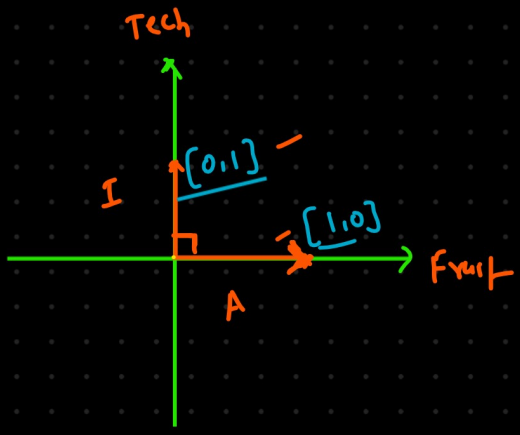
$= \sqrt{34}$.

$\vec{OP} = M \neq 0$

$= \sqrt{34} \begin{bmatrix} 5 \\ 3 \end{bmatrix} \begin{matrix} x \\ y \end{matrix}$

Vector ⇒ Direction

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



	Fruit	Tech
Apple	1	0
Iphone	0	1

Apple = [1, 0]

Iphone = [0, 1]

① Dot product ③ ecuclidean Dist } Similarity → RAG → Similarity search

② Cosine Similarity

① $[1,0]_{x_1 y_1}$, $[0,1]_{x_2 y_2}$ = $1 \times 0 + 0 \times 1 = 0$

② $\theta = 90^\circ$, $\cos 90^\circ = 0$

③ $\sqrt{(0-1)^2 + (1-0)^2} = \sqrt{2}$

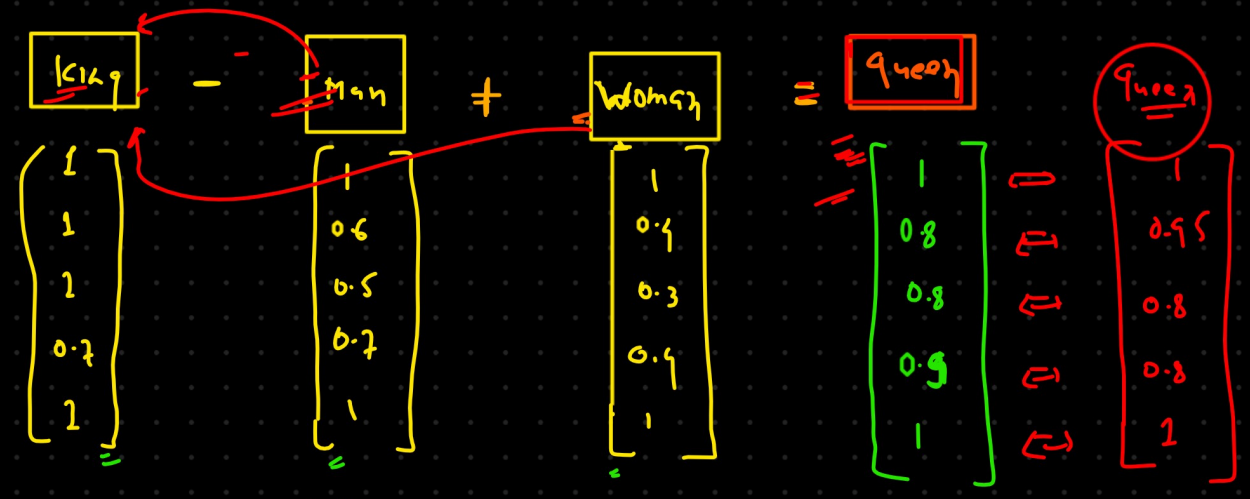
Word2vec

Scenerio

	F_1 Gender	F_2 Wealth	F_3 Power	F_4 Weight	F_5 Speak
King	1	1	1	0.7	1
queen	1	0.95	0.8	0.8	1
Man	1	0.6	0.5	0.7	1
woman	1	0.4	0.3	0.9	1
Monkey	1	0	0	0.3	0

(25) (0,1) (myself)

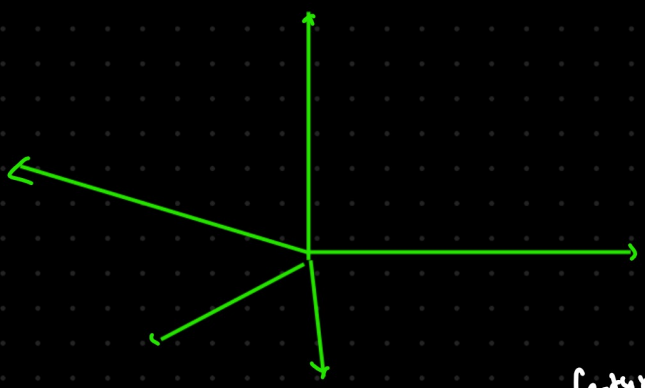
Word → Vector

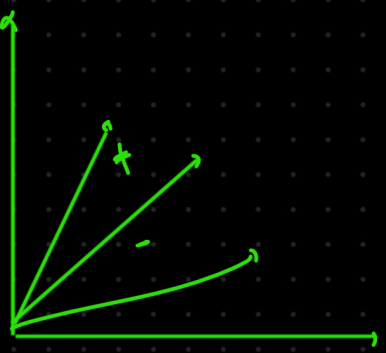


Dimension ⇒ 5D

Arithmatic operation

mathematically





1 pdf = 10 feature

5 → feature ⇒ manually

Data → feature

⇓

Automated.

10GB hours ⇒ feature ⇒ NN ⇒ Embedding

News Data

Representing

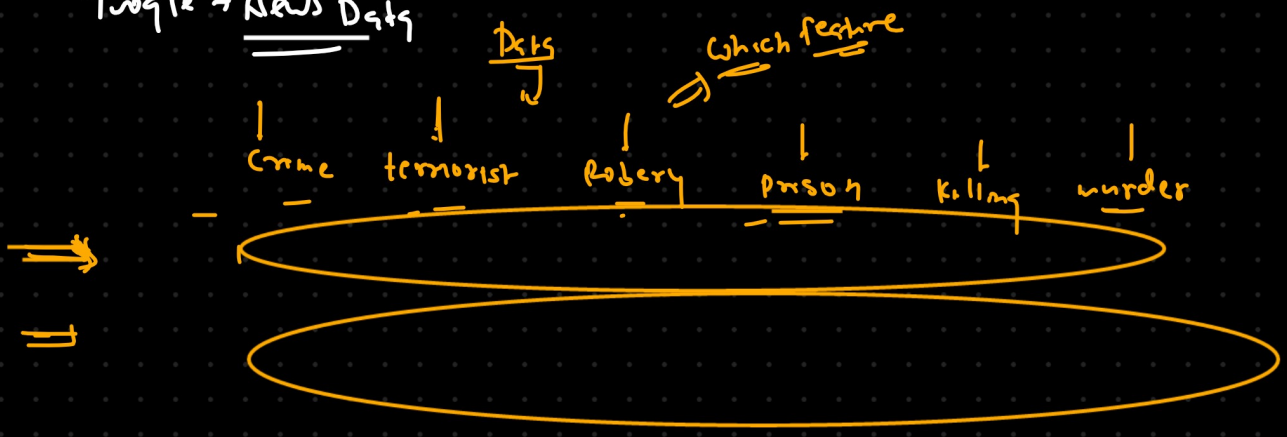
As worries about how social media sites might make people addicted grow, the European Union has recently started an investigation into Meta Platforms Inc. and their goods, like Facebook and Instagram. This move is meant to deal with and control

Object

F_1 F_2 F_3 F_4 F_5 F_6

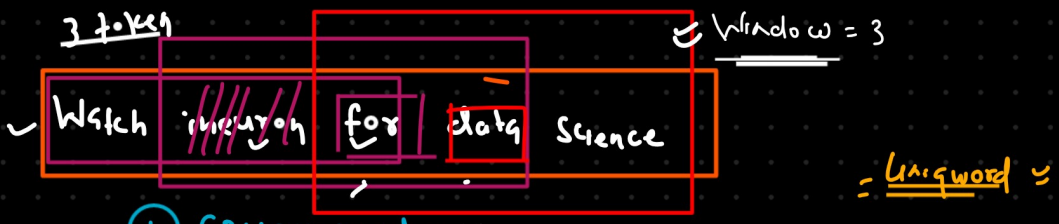
SM Addict Grow Investigate Facebook Instagram

Google → NEWS Data



Skipgram, CBOW → NN

CBOW → Continues Bag of word

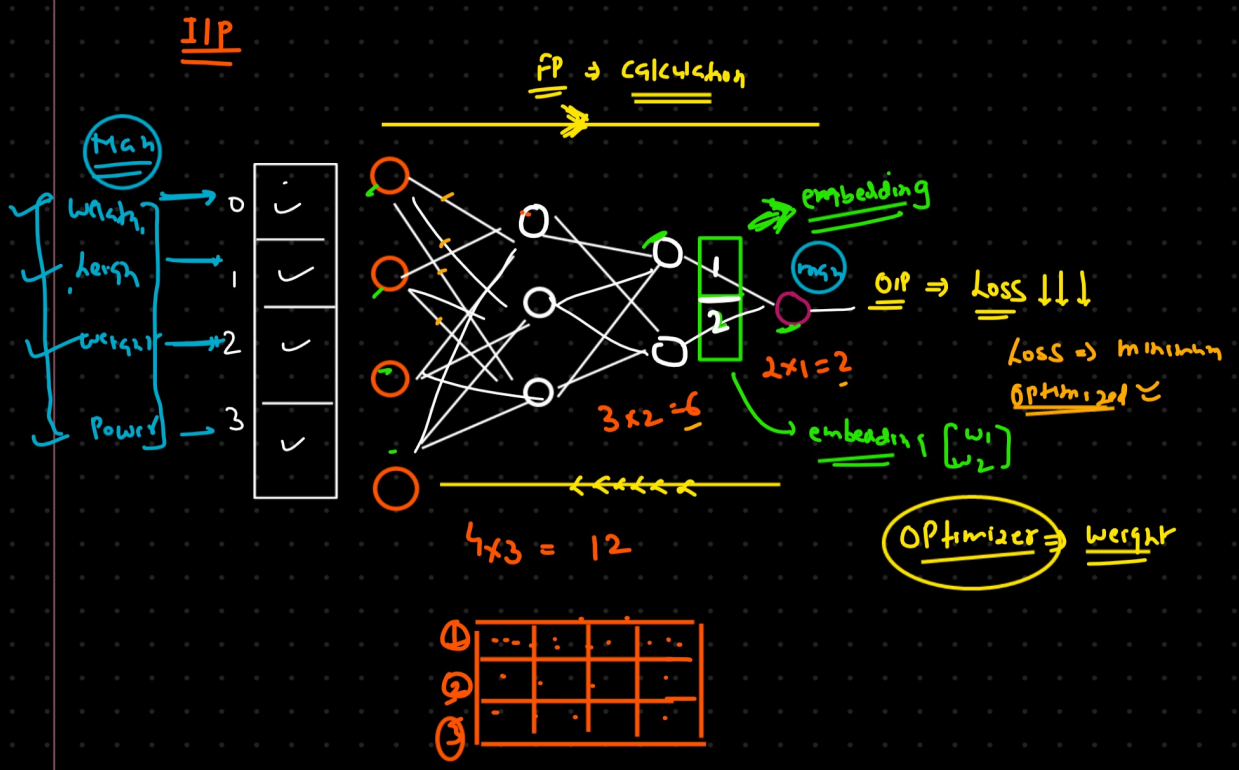


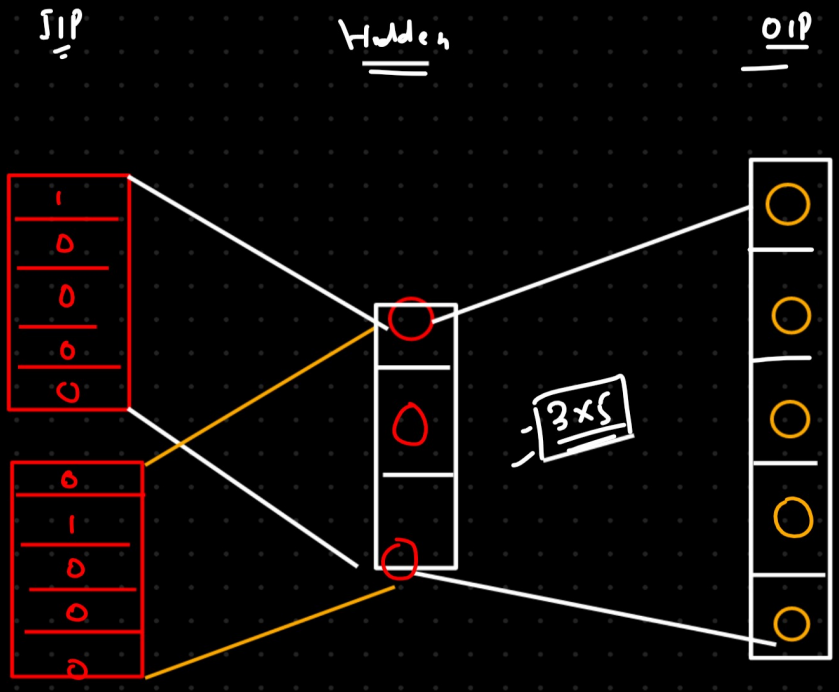
= Unigword

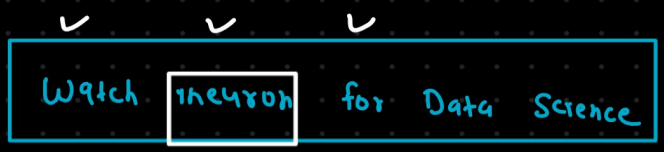
① Context window

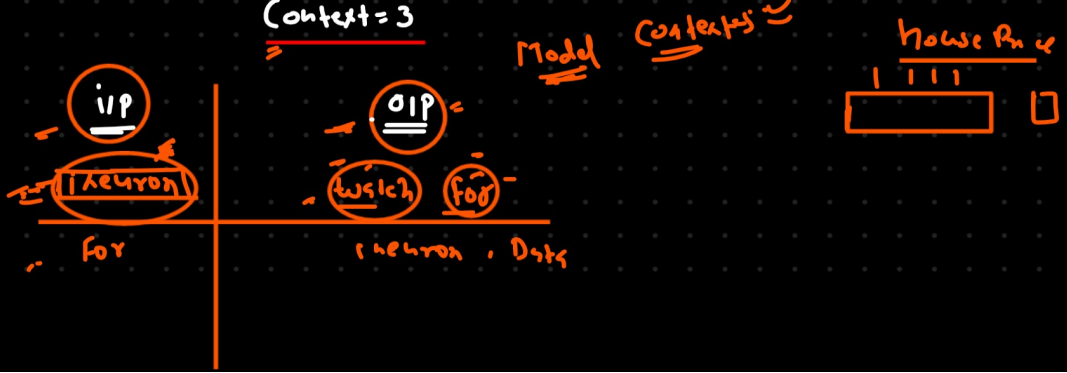
② target

X	Y
[input text length]	O/P
Watch ineuron	ineuron
ineuron data	For
For Science	Data


IIP
FP ⇒ Calculation
Mach
Watch
Learn
Weight
Power
0
1
2
3
embedding
OIP ⇒ Loss ↓↓↓
Loss ⇒ minimum
optimized
2×1=2
3×2=6
4×3 = 12
embedding [w1 w2]
Optimizer ⇒ weight


IIP
Hidden
OIP
3×5


Watch
Neuron
for
Data
Science


Context = 3
Model
Contexts
House Rule
IIP
OIP
Neuron
Watch
for
for
Neuron
Data

- Data | Fori Scenes

VoC → 5