Agenda:

① How to perform text preprocessing

→ HTML Tags remove
→ punctuation remove
→ emoji Handle
→ lamatization
→ steaming
→ lowe case

② Text Representation / word Embeddings / Text vectorization

→ One Hot Encoding
→ BOW (Bag of word)
→ TFIDF
→ Word2Vec

③ project → sentiment Analysis using ML Approach

→ ML Algo
→

Lower case

① My name is Bappy
② bappy is a Data scientist

Bappy bappy
bappy banny →

You chat on Whatsapp, Instagram, Telegram

* I want the money as soon as possible

* I want the money ASAP →

Chatbot → Amazon customer → chatbot → Mansin question

This Pizza is Amzing → Amazing

spell mistake

stopword

This pizza is very Testy I am very happy To be here in this store, Awesom exprience

positive

Increasin dim → computation

same

playing, played, plays

This movie is Amazing ☺

positive

stemy → Play

ID

☺ 😐 ♥ ⟶ possitive

Text Representation

corpus

Vocabulary

Document

Words

Natural language processing is an interdisciplinary subfield of computer science and linguistics. It is primarily concerned with giving computers the ability to support and manipulate speech.

① What is feature Extraction from text

② Why we need it?

③ Why it is so dificult

④ What is the core idea?

⑤ What are the techniques there?,

ML: tabular Data, csv, xsl

# Room	Area	Location	price
5	20	Delhi	5cr
—	—	—	—
—	—	—	—
—	—	—	—

X

Room
Area
Loc

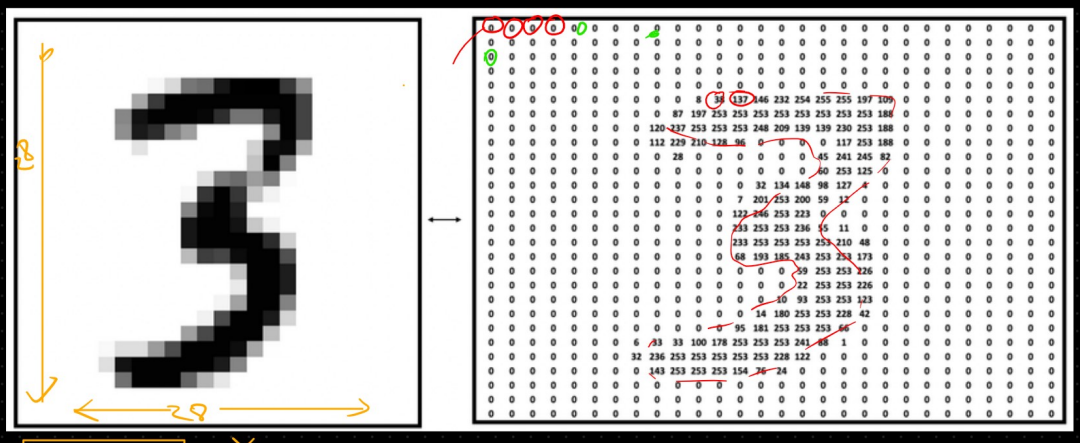
ML

Training

Y

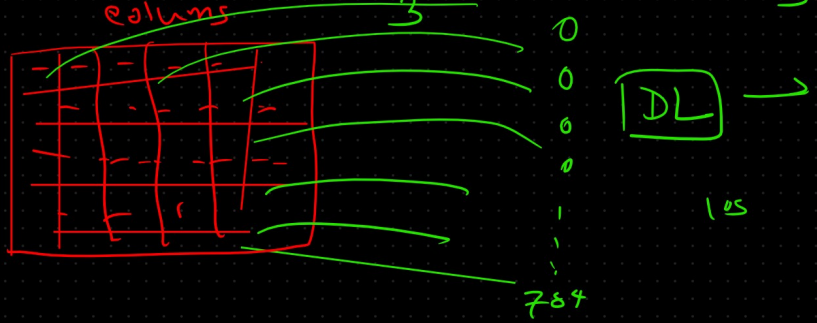
price

CV: Images, Videos,



28×28

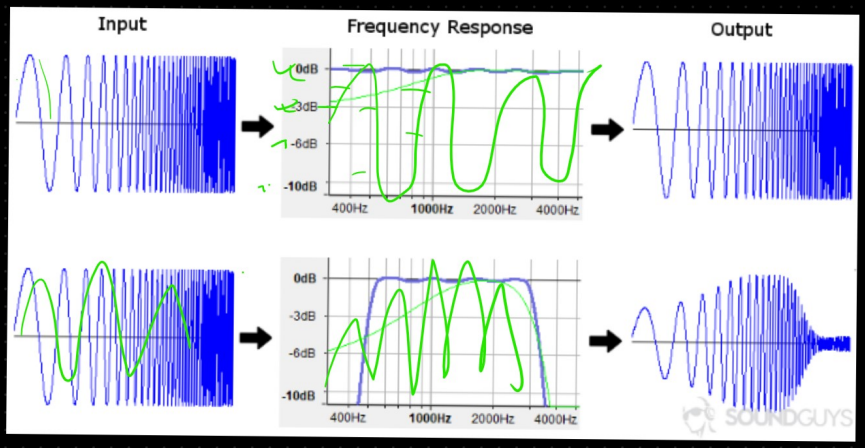
28×28
= 784



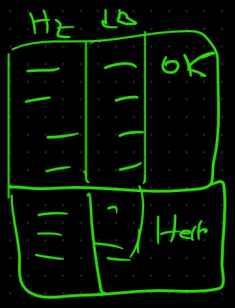
OK

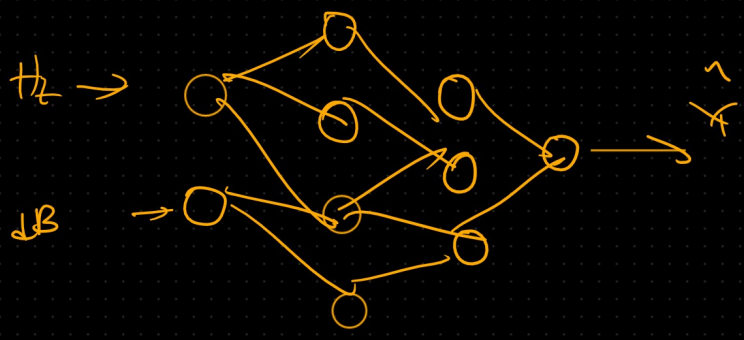
speech

Hello

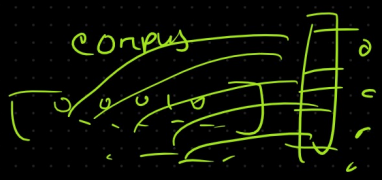


dB



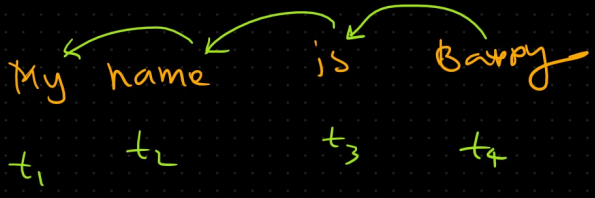


Text data in NLP



My name is Barry

0 1 2 3



One Hot:

D_1	people watch dswithbarry
D_2	dswithbarry watch dswithbarry
D_3	people write comments
D_4	dswithbarry write comment

corpus:

people watch dswithbarry dswithbarry watch dswithbarry people write comments dswithbarry write comment

One Hot:

D_1	people watch dswithbappy
D_2	dswithbappy watch dswithbappy
D_3	people write comments
D_4	dswithbappy write comment

corpus:

people watch dswithbappy dswithbappy watch dswithbappy people write comments dswithbappy write comment

$N = 5$

people	watch	dswithbappy	write	comment
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	1	0	0

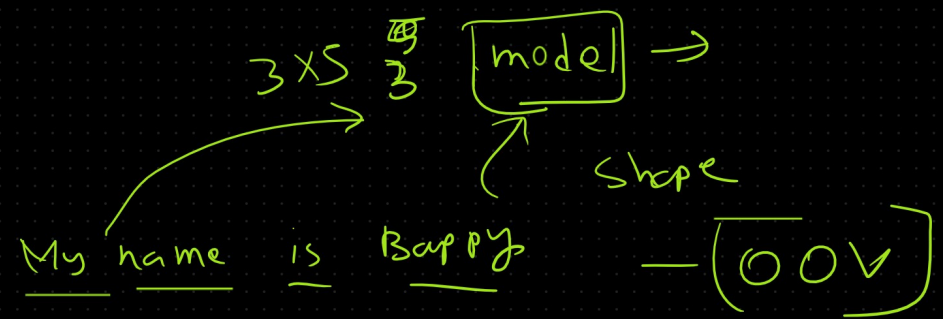
Sparse-matrix

$D_1 = [[1\ 0\ 0\ 0\ 0], [0\ 1\ 0\ 0\ 0], [0\ 0\ 1\ 0\ 0]]$

3 X 5

$D_2 = [[0\ 1\ 0\ 0\ 0], [0\ 1\ 0\ 0\ 0], [0\ 1\ 0\ 0\ 0]]$

Sent → 1000 1000



* Drawbacks:

1. Sparsity
2. No fixed size
3. OOV
4. No capturing semantic meaning

* POS: easy to Implement

Bag of words

		X	Y
1	D_1	people watch dswithbappy	1
2	D_2	dswithbappy watch dswithbappy	1
3	D_3	people write comments	0
4	D_4	dswithbappy write comment	0

people	watch	dswithbappy	write	comment
1	1	1	0	0
0	1	2	0	0
1	1	0	1	1
0	0	1	1	1

Research proved during sentiment Analysis Bag of word is best technique

This movie is WOW → positive → passive

This movie is Amazing. Amazing performace by SRK. WOW!

2

Very bad movied. bad performed by X--

Negative

2

This movie is Amazing. Amazing perform by SRK

But very bad performace by Deepika →

1

2

n-grams → 3

This movie is very good → positive

This movie is not very good → Negative