NLP Pipeline

① Data Aquisition → pre-processing → Feature Engineering / extraction

* Feature Extraction

① What is feature extraction & why do we need it?
② Why is it difficult & what is the core idea behind it?
③ What are the techniques required for this?

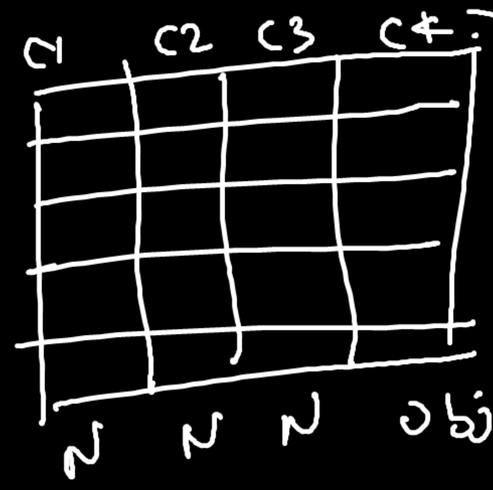
Text → Vector

ML/DL - Algorithms → maths - numbers

$y = f(x)$

NO

ML - Tabular data ✓ -

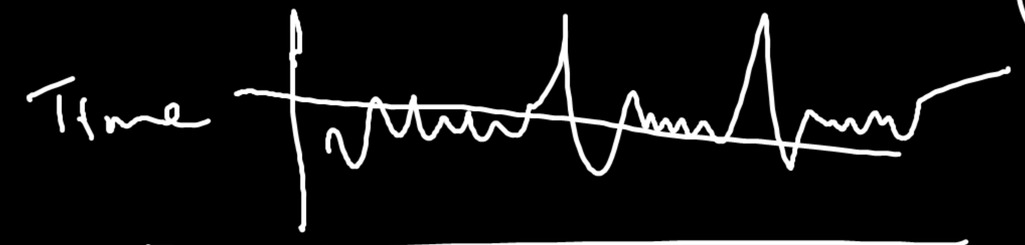


→ Significance

→ Time Series forecasting

Date & Time - Numeric/Target ✓

→ heart rate



Frequency - Numeric →

→ Text data

→ Hi, my name is Kumar Subbaram.

X my is Hi Sundar hero kumar

② Image

Pixel value

NO

NLP

Feature extraction

- OHE
- BOW
- N-Grams
- TF-Idf
- Custom Feature

Word 2 Vec embedding Glove etc.

* Common Terms

→ Corpus - consolidated Docs → Words - each Sentence/Doc has many words here

→ Vocabulary = unique words in the Documents

→ Documents - 11 / Sentence / Review

S1 I like NLP — 3

S2 NLP is in Demand — 4

S3 Learnby teaches NLP — 3

Training

S4 > NLP [redacted]

Word - S1 = 3
S2 = 4
S3 = 3

Documents = 3

Vocab = unique = 8

Corpus = 1 { I like NLP NLP is in Demand learnby teaches NLP }

I like NLP is in Demand learnby teaches

Objective : Text → Number / vector

How will we do it?

= 8 K vocab

I like NLP is in Demand Learn b

THE: [1000000] [0100000] [0010000] [0001000] [0000100] [0000010] [0000010]

Real time Problem - 50K vocab ⇒ complex

↓

Sparse matrix - More no zeros

Approach 1 = OHE

OHE

↓

only where we have some 1s

Location	D	M	P	K	C	A	
Del	1	0	0	0	0	0	(1 0 0 0 0 0)
Mus	0	1	0	0	0	0	(0 1 0 0 0 0)
Pun	0	0	1	0	0	0	(0 0 1 0 0 0)
Kut	0	0	0	1	0	0	(0 0 0 1 0 0)
Che	0	0	0	0	1	0	(0 0 0 0 1 0)
Ans	0	0	0	0	0	1	(0 0 0 0 0 1)

D1 : I like NLP [[10000000] [01000000] [00100000]]

Bag of words (BoW)

Vocab = 8

	I	like	NLP	is	in	Demand	Learnsay	teaches
D1	1	1	1	0	0	0	0	0
D2	0	0	1	1	1	1	0	0
D3	0	0	1	0	0	0	1	1

D1 :- I like NLP : [1 1 1 0 0 0 0 0]

→ This concept is called BOW idea :- Text → vector

BOW

- Binary BOW
- Count BOW

100

NLP is really in Demand

NLP & NLP concept is required

100 100

	NLP	is	really	in	Demand	and	concept	require	
D1 =	1	1	1	1	1	0	0	0	] Count BOW
D2 =	2	1	0	0	0	1	1	1	
D2 =	1	1	0	0	0	1	1	1	] Binary BOW

Pros

1. BOW is extremly simple & intuitive

Cons :-

1. Sparsity
2. OOV - Out of Vocab
   ↳ ignore the new vocab

I like Pizza → ① Remove stopwords
I don't like Pizza ②