NLP

Natural Language Processing

Pre - requsites

① Python
② stats
③ ML algo
④ ANN, optimizer

① Roadmap

why NLP?

→ Where text is data, we need the machine to understand

Roadmap

↳ ① Text Preprocessing → Tokenization, stemming, Bag of words, TF-IDF, Word to Vec, STOPWORDS, Lemmatization

② ~~Text Preprocessing 2~~ Unigrams, Bigrams

③ Text preprocessing 2 - Avg word2Vec, gensim

③ ML use cases - ChatBots, etc.

④ RNN, LSTM RNN, GRU RNN

⑤ Advanced Preprocessing - Word embedding

⑥ Bidirectional LSTM, Encoders, Decoders, Attention model

⑦ Transformers Bert.

Library

NLTK
Spacy } ML
TextBlob

Tensorflow } DL
Pytorch
Hugging Face

① Tokenization

Converting sentence into words

② stopwords :- to, of, a, an, the, etc. :

~~Step~~

③ Stemming :- Process of Reducing words to their base word stem

Disadvantage

→ The base word may not have any meaning

eg:- historical } histori   finally } → fina
history   final
finalized

Advantages

→ It is really fast

④ Lemmatization → This has dictionary of words

history } history
historical

Advantage

① Get a meaningfull word

Disadvantage

① It is slow

Usecase

Steming

↳ spam classification

↳ Review classification

Lemmation

① Text generation

② Language translation

③ chatbot

Step 2 :- Words to Vectors

① Bag of words

② TF-IDF

Terms

① Corpus → D_1, D_2, D_3, D_4 ↷ content → Para

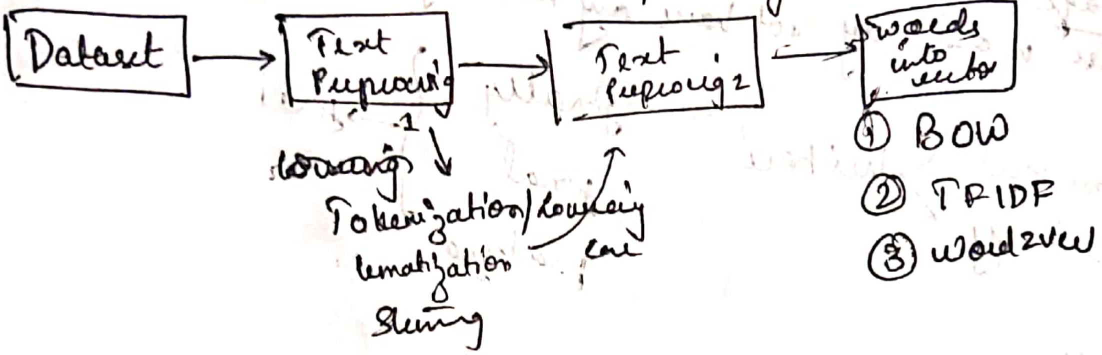
② Documents → sentence

③ Vocabulary → unique words

④ words

example

	Text	O/P
D1	The food is good	1
D2	food is bad	0
D3	Pizza is amazing	1
D4	burger is bad	0



① **One hot Encoding**

Corpus :- A man eat food

→ Cat eat food

→ People watch Krish YT

Vocabulary

A man eat Cat ppl watch krish YT

D1 [[1 0 0 0 0 0 0 0 0]

[0 1 0 0 - - - -]

[0 0 1 0 . . . .]]

D1 → [[1 0 0 0]
[0 1 0 0]
[0 0 1 0]
[0 0 0 1]] features

Issues

① Sparse Matrix

② features will change on size on sentence

③ Semantic meaning within the word is not capture

④ Out of vocabulary

Advantages

① simple to Implent

② Intuitive

② Bag of words

D1 :- He is a good boy

D2 :- She is a good girl

D3 :- Boys and girls are good.

→ Applying stopwords

D1 :- good boy

D2 :- good girl

D3 :- Boys girls good

Vocabulary	frequency
good	3
boy	2
girl.	2

→ order will be based on frequency

	d1 good	d2 boy	d3 girl.
D1	1	1	0
D2	1	0	1
D3	1	1	1

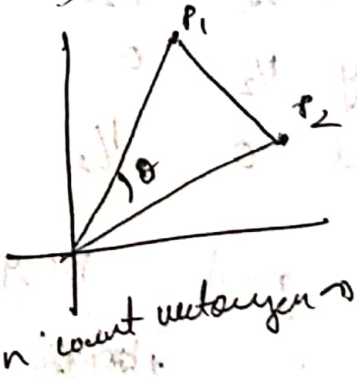
o/p

Advantages and disadvantages

- Advantages
  - → simple and intuitive
  - →
- → Sparsity
- → Out-of vocabulary
- → ordering of the words has completely changed

Cosine Similarity (Application of cosine Rule)

$$\boxed{1 - \cos\theta = \text{Cosine Similarity}}$$



In order to capture semantic info

→ Ngrams :- Bigrams, Trigrams → to follow the same order

→ In count vectorizer → hyperparam

	good (d1)	boy (d2)	girl (d3)	d4 good boy	d5 good girl
D1 :-	1	1	0	1	0
D2 :-	1	0	1	0	1
D3	1	1	1	0	0

Anish Eats food

Bigrams: Anish Eats

:- Eats food

③ TF-IDF

Term Frequency and Inverse Document Frequency

D1 :- good boy

D2 :- good girl

D3 :- boy girl good

} More weightage to rare words

$$\text{Term Frequency} = \frac{\text{no. of repetitions of words in sentence}}{\text{no. of words in sentence}}$$

$$\text{IDF} = \cancel{\log}\ \ln\left(\frac{\text{no. of sentences}}{\text{no. of sentences containing the word}}\right)$$

$$\boxed{\text{TF-IDF} = \text{T.F} \times \text{IDF}}$$

ex:-

TF → For sentence

	D1	D2	D3
good	1/2	1/2	1/3
boy	1/2	0	1/3
girl	0	1/2	1/3

IDF → for words

good	$\ln\left(\frac{3}{3}\right) \Rightarrow \underline{0}$
boy	$\ln(3/2)$
girl	$\ln(3/2)$

good — boy — girl

× ~~ln(3/2)~~

	good	boy	girl
D1	0	$\frac{1}{2}\ln 3/2$	~~$\frac{1}{2}\ln 3/2$~~ 0
D2	0	0	$1/2 \ln 3/2$
D3	0	$1/3 \ln(3/2)$	$1/3 \ln(3/2)$

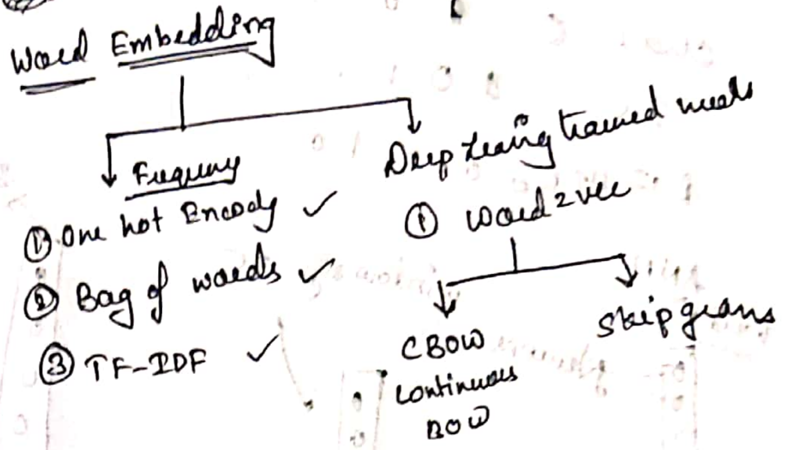
Advantages

① Intuitive
② word Importance is getting capture

Disadvantage

① Sparsity
② Out of vocabulary

① Word Embedding
② Word2vec → CBOW
→ skipgram



② **Word2vec**

Feature Representation :
① Limited dimensions
② sparcity is reduced
③ Semantic meaning is maintained

	d1 Boy	d2 girl	d3 King	d4 Queen	d5 Apple	d6 Mango
Gender			-0,92	+0.93	0	0.4
[illegible]	-1	1		0.96	-0.02	
Royal	0.01	0.02	0.95			

* **Word2vec**

① CBOW (continuous Bag of words)

CORPUS :- KRISH CHANNEL IS RELATED TO DATA SCIENCE

WINDOW SIZE :- 5

centre word is O/p, target & Content word

TRAINING DATA

Independent feature	O/P
① Krish, CHANNEL RELATED TO	IS
② CHANNEL, IS, TO, DATA	To
③ IS RELATED DATA SCIENCE	TO

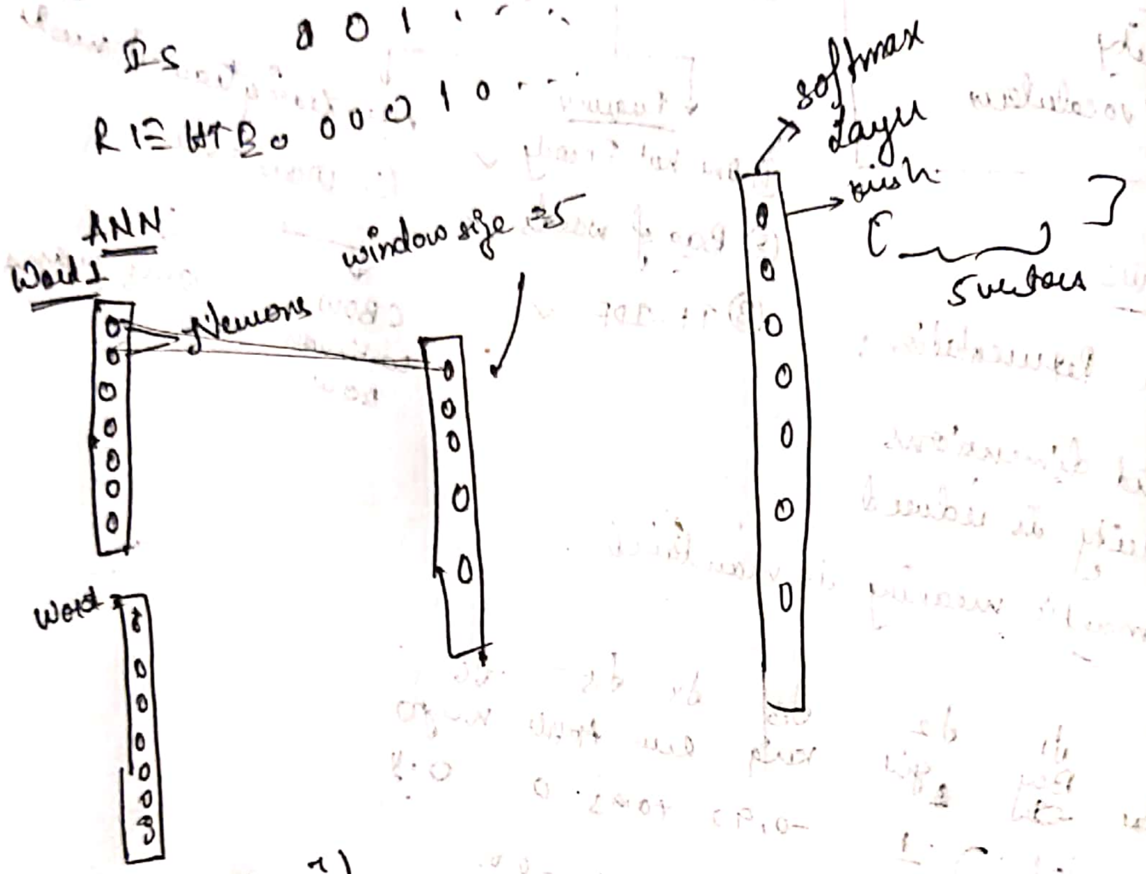
Bow. (Hence this is called continuous BoW

Krish 1 1 0 0 0 0 0 0

Chand 0 1 0 ...

Is 0 0 1 ...

RIEHTBo 0 0 0 1 0 ...



Loss = $(y-\hat{y})$

Skipgram

I/P	O/P
Is related to	other columns