Day 2 - NLP For ML And DL

Agenda

{NLTK}

① Text preprocessing → Words → Vectors

a) OHE (One hot Encoding)
b) Bag of Words (BOW)
c) TF-IDF (Term Frequency - Inverse Doument Frequency)
d) Word2Vec

a), b), c) ⇒ practical Implementation ⇒ ngrams

② Quiz → Live → 5000Rs INR

1st Prize → 2000 Rs INR
2nd Prize → 1500 Rs INR
3rd Prize → 1500Rs INR

Basic Terminologies Used In NLP

① CORPUS ✓ → Paragraph → [D1, D2, D3, D4]
② Documents ✓ → Sentence
③ Vocabulary ✓ ⇒ 10K unique words
④ Word → word

Sentiment Analysis

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

⇑
DATASET

→ Dictionary Book

⇓

10K unique words

DATASET → Text -1 Preprocessing → Text Preprocessing -2 → Words ↳ vectors

Tokenization,

① STEMMING

① BOW ?

lowering the case of words ② Lemmatization ① TFIDF ∫ ⇒

③ STOPWORDS ③ Word2Vec

① One hot Encoding

Paragraph

Vocabulary

9 ⇓ food

→ [A man eat food

→ Cat Eat food dog } → size ⇒ CORPUS

→ People watch KRISH YT]

A man eat Cat People watch KRISH YT

Out of vocabulary ⇒ CANNOT TRAIN THE MODEL

D1 → [[1 0 0 0],
[0 1 0 0],
[0 0 1 0],
[0 0 0 1]]

D2 - [[1 0 0]
[0 1 0],
[0 0 1]]

Advantages

① Simple to Implement
② Intuitive

Disadvantage

① Sparse Matrix ✓
② OOV {out of Vocabulary} ✓ [Extra Test data]
③ Not fixed size ✓
④ Semantic meaning between word is not captured

② Bag of Words

Stop words

Lower all the words case

D1 → He is a good boy

D2 → She is a good girl ~~CAT~~

D3 → Boy nd girl re good

D1 → good boy good

D2 → good girl

D3 → Boy girl good

Vocabulary	Frequency
good	3
boy	2
girl	2

	f1	f2	f3
→	good	boy	girl
– Doc 1 →	1	1	0
– Doc 2	1	0	1
Doc 3 →	1	1	1

O/p ⇒ Assumptions

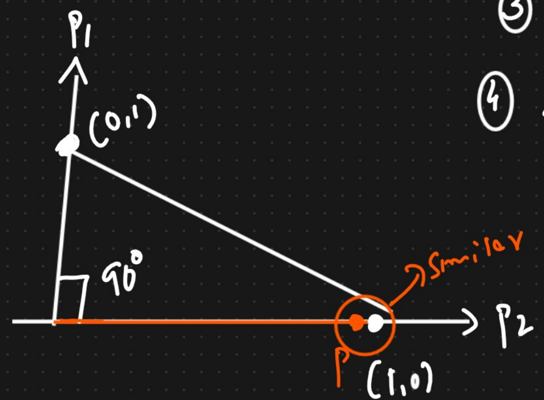
{ Euclidian Distance
Cosine Similarity. }

BoW ⇒ Binary BoW

Advantages

① Simple and Intuitive

Cosine Similarity



Disadvantages

① Sparsity

② OOV

③ Ordering of the words.

④ Semantic meaning Not able to
Cos 45 = 0.53 capture }.

$1 - 0.53 = \text{Cos-similarity}$

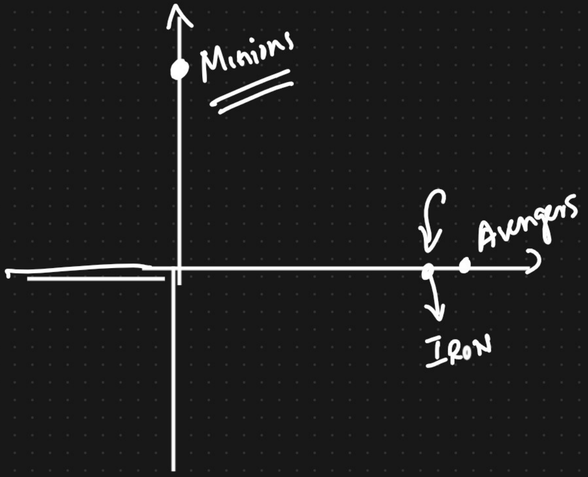
0.47

Cos 90 = 0

$1 - 0 = \boxed{1}$

Cos 0 = 1

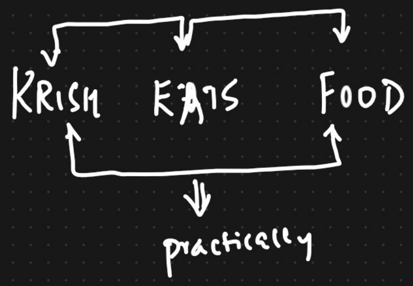
$1 - 1 = 0$



Capture the semantic Info

Ngrams ⇒ Bigrams, Trigrams, — N grams

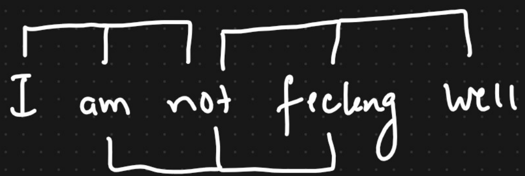
	f1	f2	f3	f4	f5
	good	boy	girl	good boy	good girl
Sent 1	1	1	0	1	0
Sent 2	1	0	1	0	1
Sent 3	1	1	1	0	0



BI-GRAMS ? 2 Bigrams

KRISH EATS   EATS FOOD

TRIGRAMS ⇒ 3 Trigrams



I am not   Am Not feeling   Not feeling well

ngrams

(1, 3)

KRISH IS NOT FEELING WELL

f1	f2	f3	f4	FS
KRISH	IS	NOT	FEELING	WELL

KRISH IS