Natural Language Processing Machine Learning

Why NLP?

I/p features → {Numbers, categories}

I/p features → Text, Sentences

Text → FORMAT → Vectors → Machine

Vectors → Numbers

Human Being

Vectors → Meaningful

① BOW
② TF-IDF
③ Word2Vec {Deep Learning Technique}

Words → Vectors.

① Corpus → Paragraph
② Documents → Sentences, Datapoints
③ Vocabulary → Unique words in this corpus
④ Words → Independent words

Dataset

	Text	O/P
D1		1
D2		0
D3		1
D4		
D5		

Words → Vectors

DATASET → Text Preprocessing 1 → Text Preprocessing 2 → Words to Vector → ML Models

Cleaning our text data.

① Stopwords
② Stemming ✓
③ Lemmatization ✓
④ Tokenization

① BOW
② TF IDF
③ N Grams
④ Word2Vec

Stemming And Lemmatization

{ PROCESS OF REDUCING THE WORD TO THEIR ROOT FORM } ⇓ Stemming

history, historical → Stemming ↓ histori ⇒ Root word

finally, final, finalized → final

going, goes, gone → go ← meaning for word.

Meaning of the word may change

happy → happi

Lemmatization

Root Form of the Word → Meaningful word.

Stop words : NLTK [⇓ Stopwords] ⇓ Sentiment analysis

the, he, she, it, if, of, an, ⇒ Not

Input size fixed

ML & DL → I/p size fixed

① One hot Encoding

D1 [A (x) man eat food]

D2 [Cat eat food]

D3 [People like Datascience] ++

⇒ Train ⇒ Corpus ⇒ stop words ⇒ Model

D3 = [[———] 4*7 [———] [———] [———]]

→ DL [KRISH YT CHANNEL] ← Test Data

Vocabulary

[man, eat, cat, food, people, like, data Science]

D1 [[1 0 0 0 0 0 0] [0 0 0 0 0 0 0]
[0 1 0 0 0 0 0] [0 0 0 0 0 0 0]
[0 0 0 1 0 0 0]] 3 × 7 ⇒ One Hot Encoding

Advantages

① Simple to Implement

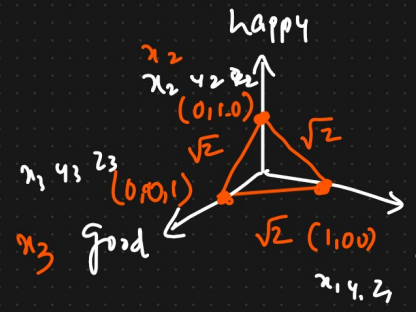
Disadvantages

① Sparse Matrix {Overfitting}
② OOV {out of vocabulary}
③ No capturing of Semantic? {Semantic meaning}
④ Input size

[angry, happy, good]

1	0	0
0	1	0
0	0	1

Opposite words { angry, happy, good }



(0,1,0) — (1,0,0) { No }

(0,0,1) — (1,0,0)

√2

$$\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$$

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

$$= \sqrt{1+1} = \sqrt{2}$$

I/p size → fixed

Test data → [Good KRISH YT] ← New words

1 0 0

② BOW {Bag Of Words} [Sentiment Analysis, Text Classification].

don't → stopwords

① D1 → He is a good boy
D2 → She is a good girl
D3 → Boy and girl are good

⇒ ① Lower the words ② Stopwords ③ Stemming

D1 → good boy boy
D2 → good girl
D3 → boy girl good

{Binary = True}

Vocabulary | Frequency

Vocabulary	Frequency
good	3
boy	2
girl	2

BOW

Vocabulary

	[good	boy	girl]	O/P
D1	1	1	0	—
D2	1	0	1	—
D3	1	1	1	—

I/P ⇒ fixed

Advantages

① Simple & Intuition
② Input fixed issue is resolved

Disadvantages

① Sparsity {Exists} ⇓ Reduced — It is less when compared to ONE
② OOV → Exists → we should ignore the new words.
③ Semantic Relationship.
Captured | Not Captured.

I like pizza
I don't like pizza ←

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

Vocab	freq
like	2
pizza	2
don't	1

⇒ ORDERING.

(1,2) ⇒ Bigram along with unigram

don't pizza.

		like	pizza	dont	like pizza	like don't	pizza don't	don't like	pizza like
D1		1	1	0	1	0	0	0	0
D2		1	1	1	1	0	0	1	0

④ N-Grams

(1,2)
very
He is intelligent boy
He is not a

Remember

1. Unigram ⇒ Bag of words
2. Bigram ⇒ Bag of words + Bigram.
3. Tri Gram
4. Quad Gram

!

n gram

(1,2)

ORDERING

KRISH EATS FOOD

FOOD EATS

KRISH EATS FOOD | KRISH EATS | EATS FOOD | KRISH FOOD | FOOD KRISH

③ Term Frequency → Inverse Document Frequency (TF-IDF)

Sent1 : good boy

Sent2 : good girl

Sent3 : boy girl good

Term Frequency : No. of rep of words in sentence / No. of words in sentence

$$IDF = \log_e \left(\frac{\text{No. of sentences}}{\text{No. of sentences containing the word}} \right)$$

Sent = Document

Term Frequency × Inverse Document Frequency

IDF

	Sent1	Sent2	Sent3	words	
good	1/2	1/2	1/3	good	$\log_e(3/3) = 0$
boy	1/2	0	1/3	boy	$\log_e(3/2)$
girl	0	1/2	1/3.	girl	$\log_e(3/2)$

TF-IDF

	f1 good	f2 boy	f3 girl	
Sent1	0	$\frac{1}{2} \times \log_e(3/2)$	0	[good boy]
Sent2	0	0	$\frac{1}{2} \log_e(3/2)$	[good girl]
Sent3	0	$\frac{1}{3} \log_e(3/2)$	$\frac{1}{3} \log_e(3/2)$	[good boy girl]

① Advantages

① Semantic meaning is well captured to some extent

② Intuitive.

Disadvantage

① Sparsity

② Out of vocabulary

③ Ordering (N Grams).

④ Computation is high.
↳ complexity.