Day 3, 13th Dec 23

NLP - by mentor Dr. Ayan Debnath, IIT Delhi + Harvard University Alumni

Recap

① Stemming → sentiment analysis → base word may/may not have meaning

② Lemmatization → root base & always have meaning → chatbot

③ Code

Topic → text → No / vectors (Numerical representation)

i/p: I love NLP → o/p: +ve (1)

21 | 236 | 75 → model → o/p (1 or 0)

Bag of Words (BOWs)

binary BOW

Sent1: He is intelligent boy
Sent2: She is intelligent girl
Sent3: Both of them are intelligent boy & girl

→ lower → stopwords

Sent1: intelligent boy & good boy
Sent2: intelligent girl
Sent3: intelligent boy & girl

text Processing → ①

→ $ He is ! intelligent boy !#

step-1: ⇓ data cleaning

He is intelligent boy

step-2: ⇓ Lowering

he is intelligent boy

step-3: ⇓ remove stopwords

intelligent boy

step-4: ⇓ Stemming / Lemmatization

step-5: Tokenization

token ID

He: 19 | 200 | 26 | 54
he

Vocabulary → 20K words

Dictionary: 1 A ... 20K z

He: 256 | love: 512 | NLP: 1000

Token ID

Words	Frequency
intelligent	3
boy	2
girl	2

Python

BoW

step ① Data cleaning
② Lowering
③ Remove stopwords
④ Stemming
⑤ Calculate Histogram

Feature set / Feature Engineering

Sl no	intelligent f_1	boy f_2	girl f_3	y
Sent1	1	1	0	
Sent2	1	0	1	
Sent3	1	1	1	

Keyword
Import info retrieve

→ Training Data

Disadvantage

① semantic info ✗
② context ✗
③ weightage ✗
④ Bigger dataset → BOW → doesn't work well
⑤ Sparsity

advantage

① Easy to implement
② Processing time ↓
③ works well with small dataset

Sent1
Sent2
...
Sent 10,000

unique words

ML
Ayan
unique words
1
10,00,000

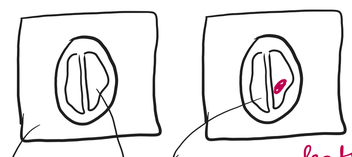
He love NLP & he also love ML

256	512	1000	25	256	19	512

text → model → ???

Ayan loves NLP

??	512



features: f_1 f_2 f_3 ... f_4 → independent Variable

dependent → y / output

Pt no.	BMR	Age	mean	median	f_4	y / output
1						NT (0)
2						T (1)
...						
10,000						

→ Training Dataset