Natural Language Processing (NLP) - Day2

Recap

1) Application - Alexa, Voice assistant, Text Summarization
2) Roadmap -
3) Data Pre-processing - cleaning
4) Stemming, Lemmatization, Stopwords
5) Hands On

BOW Bag of words

text X ⟶ No.

boy girl

Sent1 : He is intelligent boy ⟶ intelligent boy

Sent2 : she is intelligent girl → intelligent girl

Sent3 : Both of them are intelligent boy & girl

word	Freq
boy	2
girl	2
intelligent	3

vector ⇒ No.s

	words f1 boy	f2 girl	f3 intelligent	o/p
sent1	2	0	1	1
2	0	1	1	0
3	1	1	1	1

features

model → = +1 / 0 → predict

Food is good → +ve / +1
" " bad → -ve / 0

Disadvantage

1) Semantic information X
2) Context X
3) Huge dataset X

	features HR	BMR	H	wt	...	D/ND
sl 1						
sl 2						D →
sl 3						
sl 4						

independent variables

$y = f(x) = f(x_1, x_2, x_3, \dots)$

TF-IDF

	intelligent words 1	w2	w3 boy
sent1	0.9 ✓		0.01
2			
3	0		

$\frac{1}{2} \times \log(3/2) =$ ✓

TF (term Frequency)

IDF ✓

$TF = \frac{\text{No. of req words in sent}}{\text{Total no. of word.}}$

$IDF = \log\left(\frac{\text{No. of sentences}}{\text{No. of " containing that word}}\right)$

words	IDF
boy	log(3/2)
girl	log(3/2)
intelligent	log(3/3)

TF

	sent1	sent2	sent3
boy	1/2	0	1/3
girl	0	1/2	1/3
intelligent	1/3	1/3	1/3

words	Freq
boy	2
girl	2
intelligent	3

	word1	w2	w3
sent1	1 → 0.9	0	1
sent 2	0	1	1 → 0.01
sent 3	1	1	1

max. features = 3000 ~~30000~~

w1 w2 .. w1000 w2000 = =

100
1000
50,000

✓ ✓ ✓