

SpK - Varsitate
INWU

Vectorization Techniques ~~OOV~~ (SOTA)

Problem	BoW & TF-IDF	W2V, GloVe (FastText)	GPT, BERT
Dimensionality & sparsity	High ✓ sparsity	Low (AKA Embedding) X Dense (50-300)	Low (AKA Embedding)
Semantic	X	✓ slow ANN	✓
Sequential information	X	X CBOW skip-gram	✓
OOV	X	X	✓
Contextual understanding	X	X	✓

B1-gram

R1 I like food | 1 do | but like | 1 0
 R2 I liked food | 1 0
 R3 I do not like food
 R4 food is good
 R5 food was okay.

vocab: 1100

[0100]

	I	like	food	do	not	good	was	okay	is
R1	1	1	1	0	0	0	0	0	0
R2	1	1	1	1	1	0	0	0	0
R3	0	0	1	0	0	1	0	0	0
R4	0	0	1	0	0	0	1	1	0

Webinar Chat

CA tuning

MY Mohamed Yunus to Hosts and panelists
reduce the common words sir

US Umil Shah to Hosts and panelists
here

we need to retain context

MY Mohamed Yunus to Hosts and panelists
We can go to phase to see the semantic meaning

SP Santoshkumar Pa... to Hosts and panelists
stemming

US Umil Shah to Hosts and panelists
yes

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...

I went to a river
 taking a bath
 the bank to disen
 loan application in

Phase I :-

Basic language representation

Problem

→ memory consumption, complexity, curse of dimensionality, interpretability

Solⁿ :-

Remove the stopwords, lower case, convert to vector form, remove the singular & plural

② → sequence information is lost

Solⁿ :-

Webinar Chat

CA tuning

MY Mohamed Yunus to Hosts and panelists
reduce the common words sir

US Urnil Shah to Hosts and panelists
here

we need to retain context

MY Mohamed Yunus to Hosts and panelists
We can go to phase to see the semantic meaning

SP Santoshkumar Pa... to Hosts and panelists
stemming

US Urnil Shah to Hosts and panelists
yes

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...

Training

$R1 = 1$ like food

$R2 = 1$ do not like food

Predict

\Rightarrow I like meal

OOV

vocab :- Bi-grams $n=2$

$I =$
 $\begin{matrix} \text{do not} & \text{like} & \text{meal} \\ \text{do not} & \text{like} & \text{meal} \end{matrix}$

Stop words

$R1 = [110]$

$R2 = [011]$

$R1 \neq R2$

	I like	like food	do not	do not like
$R1$ <u>like</u>	1	1	0	0
$R2$ <u>(-)ve</u>	0	1	1	1

I like food do not

is

n-gram = (1,1,1)

I like food do not

I like like food

Food do do not

I like food

food do not

I like food do

like food do not

I like food

Webinar Chat

Urnil Shah to Hosts and panelists

US yes it can

SK Sourav K to Hosts and panelists

SK yes

US Urnil Shah to Hosts and panelists

US sure sir

R rashmi to Hosts and panelists

R yes pls

US Urnil Shah to Hosts and panelists

US n-gram cannot identify significant words, is that right

N Nagarajan to Hosts and panelists

N OOV is solved by using n-grams sir

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...