Day 3 - NLP

Agenda

① BOW {Bag of words}

② Tf Idf

③ Practical Implementation

④ Quiz

Spam Classifier

Word2vec

① Embedding Layer

② Word2vec → CBOW ✓ → Skip gram ✓

③ Architecture

④ Practical Problem

⑤ Glove

Bag of Words → Text → Vectors

①

Good

Sent 1 → He is a good boy

Sent 2 → She is a good girl

Sent3 → Boy and girl are good

Stopwords ⇒ lowering

②

Sent1 good boy

Sent 2 good girl ~~Ca~~4

Sent 3 boy girl good

③ Frequency (Vocabulary)

	frequency
good	3
boy	2
girl	2

④

$$\begin{array}{c|ccc} & \underset{f_1}{\text{good}} & \underset{f_2}{\text{boy}} & \underset{f_3}{\text{girl}} \\ \hline \text{Sent 1} & 1 & 1 & 0 \\ \text{Sent 2} & 1✓ & 0 & 1✓ \\ \text{Sent3} & 1 & 1 & 1 \end{array}$$

Not all Similar

opposite

The food is good

The food is not good

$$\begin{array}{ccccc} \text{food} & \text{good} & \cancel{\text{is}} & \text{not} & \cancel{\text{The}} \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{array}$$

⇓

Similar

⇓

{ TF-IDF }

Term Frequency – Inverse Document Frequency   TF - IDF

Sent 1 : good boy ✓

Sent 2 : good girl ✓

Sent 3 : boy girl good ✓

$$\text{Term Frequency} = \frac{\text{No.of rep of words in sentence}}{\text{No.of words in sentence}}$$

⇓ Sentences

X

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

⇓ words

Term Frequency * Inverse Document Frequency

	Sent1	Sent 2	Sent3
good	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$
boy	$\frac{1}{2}$	0	$\frac{1}{3}$
girl	0	$\frac{1}{2}$	$\frac{1}{3}$

*

Words	IDF
good	$\log_e(3/3) = 0$
boy	$\log_e(3/2)$
girl	$\log_e(3/2)$

⇓

	f1 good	f2 boy	f3 girl
Sent 1	0 ✓	$\frac{1}{2} \times \log_e(3/2)$ ✓	0
Sent 2	0 ✓	0	$\frac{1}{2} \times \log_e(3/2)$ ✓
Sent 3	0 ✓	$\frac{1}{3} (\log_e(3/2))$ ✓	$\frac{1}{3} \log_e(3/2)$ ✓

O/p

Advantages

① Intuitive

② Word Importance is getting capture

Disadvantage

① Sparsity

② Out of vocabulary

good ✓	10
boy ✓	7
girls ✓	6
toy	5
|	1
.	1
.	1
!	2
	<3

max.f = 3

good boy girls

[ ]

Assignment

Take any Text Datset

{ Appy BOW, Tf IDF, hgram } ✓ ✓

Krishnaik06@gmail.com