Natural language processing (NLP)

- Jaya Sicora Mahiswara Rahman Chen

by Dr Rajeev Dabhade, Data Science Mentor
IIT Delhi Alumni, Harvard Alumni

Recap: → Bag of words
→ TF-IDF

⇒ sequence ...

	w1	w2	w3
s1	1	0	0
s2			

→ model (ML, RF)
↳ Additional

Disadvantages

① semantic information is deteriorated
↳ meaning/context/interpretation
② chance of overfitting

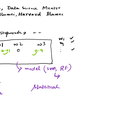
Word2Vec:

word → no. (from matrix (BOW))

200
no → 0 0 0 1 0 1 0

→ each word is represented as vector
→ semantic information
→ relationship b/w words

n/10,000

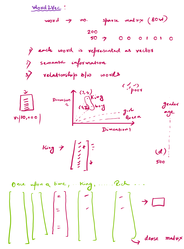


King → $\begin{bmatrix} 0.9 \\ 0.7 \\ \vdots \end{bmatrix}$

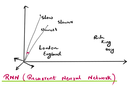
(d) 300

King - Boy = $\begin{bmatrix} 0.2 \\ 0.1 \\ \vdots \end{bmatrix}$ 300

Man → 0/1
↳ 0.9
↳ $\begin{bmatrix} 0.1 \\ 0.2 \\ \vdots \\ 0.6 \end{bmatrix}$ 32



Embeddings



RNN (Recurrent Neural Network)

I love NLP. I am working on chatbot ...
s_1 s_2 s_3

y_hat →

I love NLP
t=1 t=2 t=3 ....

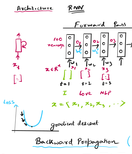
Application

① Next word prediction
② Auto-correction
③ Auto completion
④ Text summarization
⑤ Language Translation
⑥ Image captioning
⑦ Speech Recognition

Advantage not ANN/FFNN

Sequence info.

Architecture **RNN**


Forward Pass
no
embedding
300
300
300