GDG DEVFEST KOLKATA

Using wordvectors

Priyanka Sinha

August 4, 2019
Novotel Kolkata Hotel and Residences

GDG Kolkata

WTM : Kolkata :

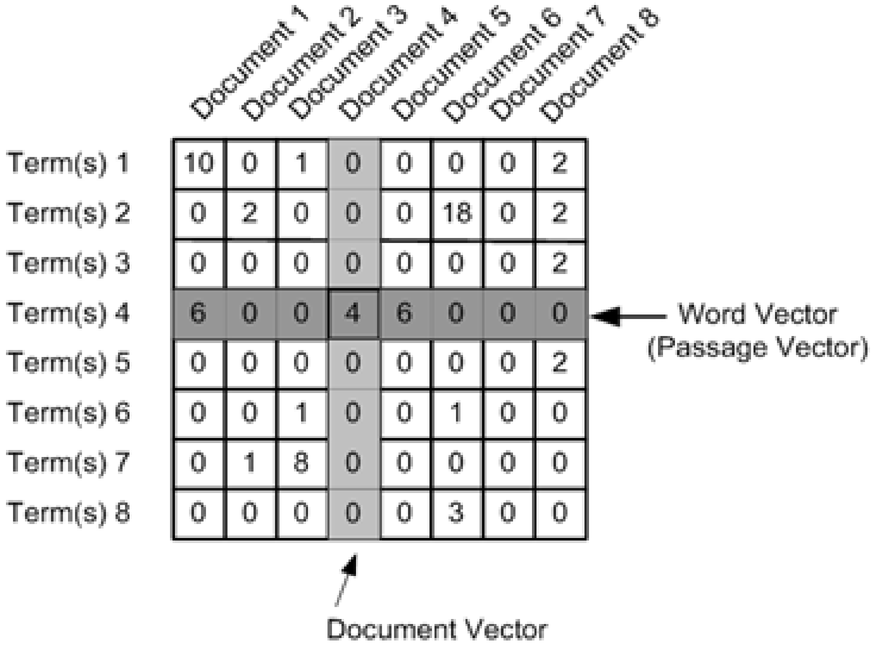
GDG Cloud Kolkata

Agenda

- Why language representation
  - Language modeling
    - Understanding meaning from words used computationally
    - Words that are used and occur in the same contexts tend to purport similar meanings
    - “a word is characterized by the company it keeps” by Firth
- word2vec
- xlnet

What are wordvectors

- one hot encoding
- Term document matrix
- Sparse representation
- Not capture semantic sense



	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

What are wordvectors

- tf-idf

sentence 1	earth is the third planet from the sun
sentence 2	Jupiter is the largest planet

Word	TF (Sentence 1)	TF (Sentence 2)	IDF	TF*IDF (sentence 1)	TF*IDF (Sentence 2)
earth	1/8	0	log(2/1)=0	0.0375	0
is	1/8	1/5	log(2/2)=0	0	0
the	2/8	1/5	log(2/2)=0	0	0
third	1/8	0	log(2/1)=0.3	0.0375	0
planet	1/8	1/5	log(2/2)=0	0	0
from	0	0	log(2/1)=0.3	0	0
sun	1/8	0	log(2/1)=0.3	0.0375	0
largest	0	1/5	log(2/1)=0.3	0	0.06
Jupiter	0	1/5	log(2/1)=0.3	0	0.06

What are wordvectors

- tf-idf
-

TFIDF

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Latent semantic indexing

- Term-document matrices are very large (millions of words)
- But the number of topics that people talk about is small (in some sense)
  - Clothes, movies, politics, ...
- Can we represent the term-document space by a lower dimensional latent space?

Singular Value Decomposition

For an M $\cdot$ N matrix $\mathbf{A}$ of rank r there exists a factorization (Singular Value Decomposition = **SVD**) as follows:

$$A = U\Sigma V^T$$

$\boxed{M{\cdot}M}$ $\boxed{M{\cdot}N}$ $\boxed{V \text{ is } N{\cdot}N}$

Singular Value Decomposition

$$A = U\Sigma V^T$$

M·M M·N V is N·N

- $AA^T = Q\Lambda Q^T$
- $AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = (U\Sigma V^T)(V\Sigma U^T) = U\Sigma^2 U^T$

The columns of **U** are orthogonal eigenvectors of $\mathbf{AA^T}$.

The columns of **V** are orthogonal eigenvectors of $\mathbf{A^TA}$.

Eigenvalues $\lambda_1 \ldots \lambda_r$ of $\mathbf{AA^T}$ are the eigenvalues of $\mathbf{A^TA}$.

$$\sigma_i = \sqrt{\lambda_i}$$

$$\Sigma = diag(\sigma_1 \ldots \sigma_r)$$

Singular values

- A simple example term-document matrix (binary)

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

- Example of C = UΣVT: The matrix Σ

Σ	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	1.28	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	0.39

LSA Example: Reducing the dimension

U	1	2	3	4	5
ship	−0.44	−0.30	0.00	0.00	0.00
boat	−0.13	−0.33	0.00	0.00	0.00
ocean	−0.48	−0.51	0.00	0.00	0.00
wood	−0.70	0.35	0.00	0.00	0.00
tree	−0.26	0.65	0.00	0.00	0.00

Σ_2	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00

V^T	d_1	d_2	d_3	d_4	d_5	d_6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00

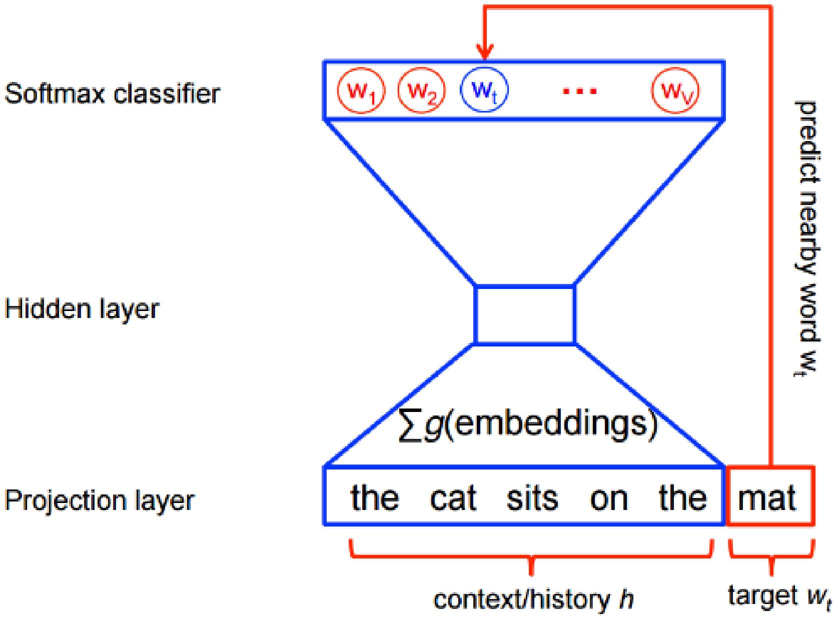
Original matrix C vs. reduced $C_2 = U\Sigma_2 V^T$

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

C_2	d_1	d_2	d_3	d_4	d_5	d_6
ship	0.85	0.52	0.28	0.13	0.21	−0.08
boat	0.36	0.36	0.16	−0.20	−0.02	−0.18
ocean	1.01	0.72	0.36	−0.04	0.16	−0.21
wood	0.97	0.12	0.20	1.03	0.62	0.41
tree	0.12	−0.39	−0.08	0.90	0.41	0.49

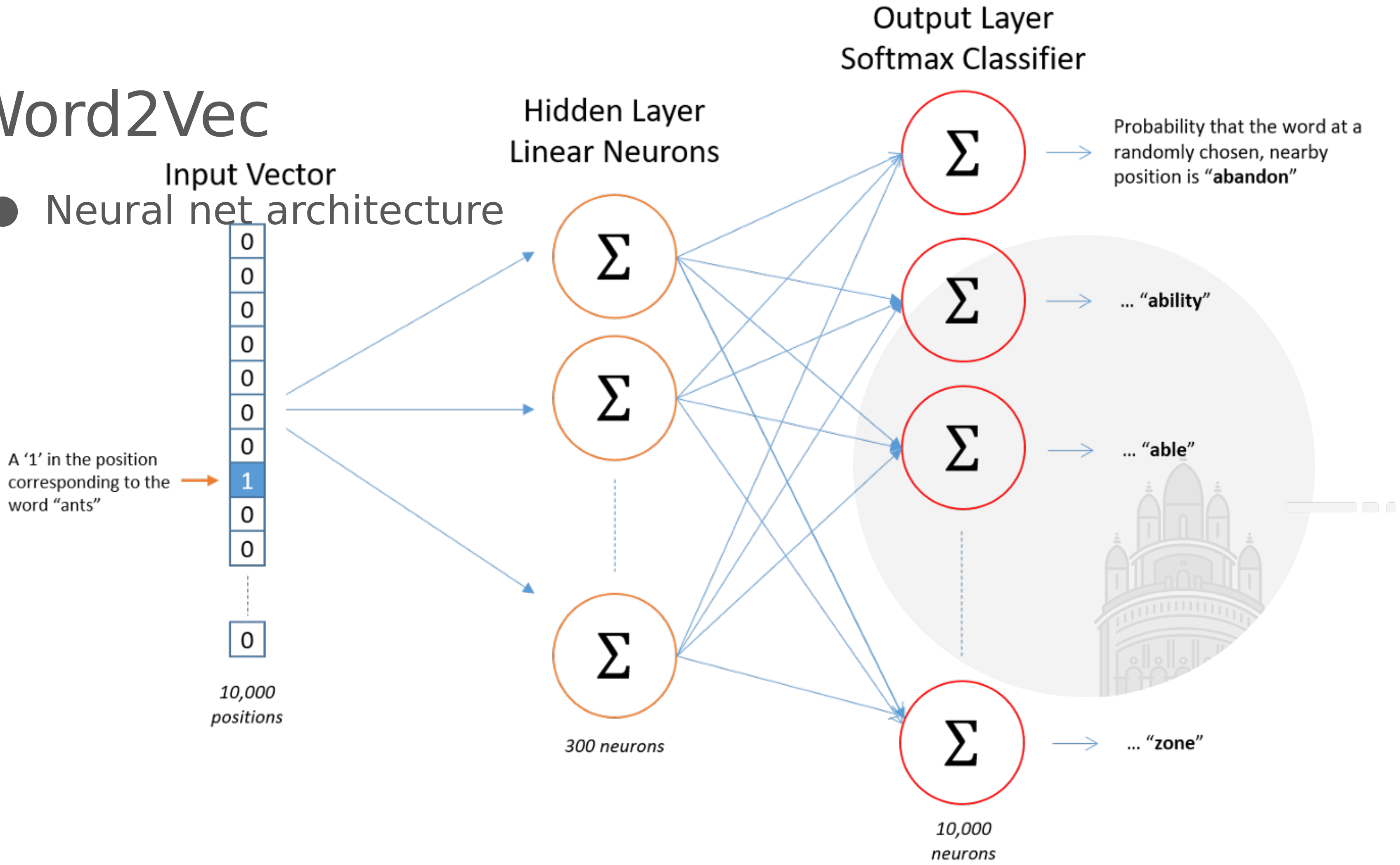
Word2Vec

- Semantic and syntactic multidimensional representation
-



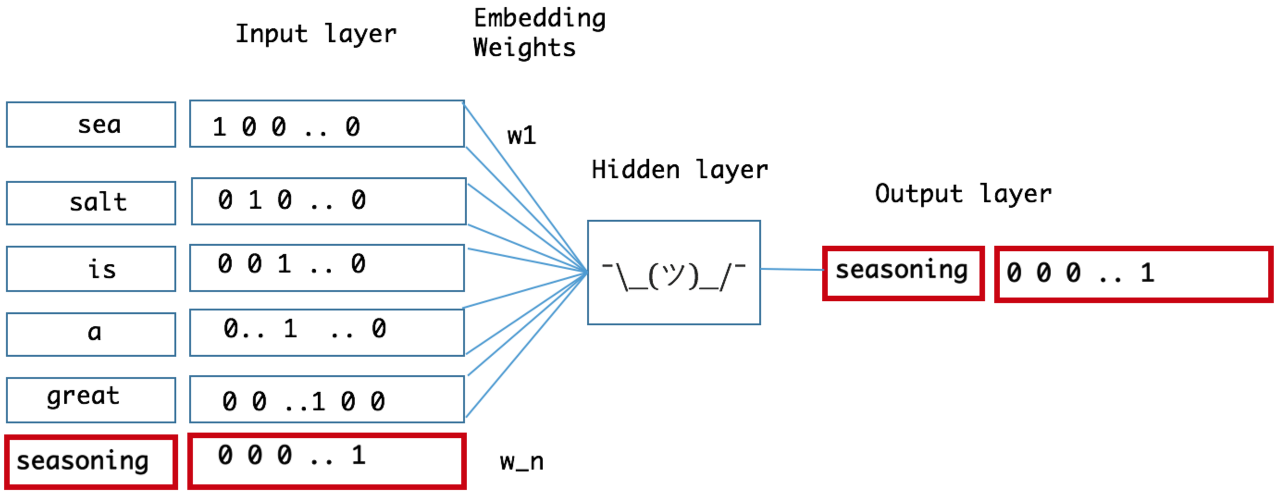
Word2Vec

- Neural net architecture

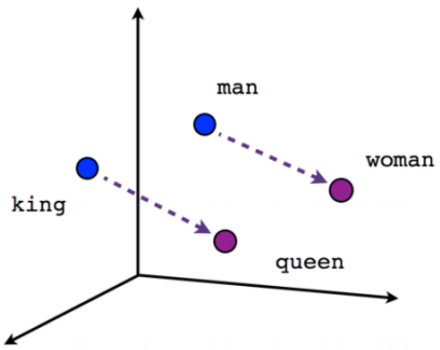


Word2Vec

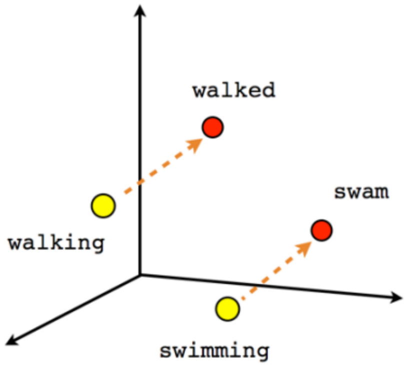
- Neural net architecture
- Word analogy examples semantic similarity



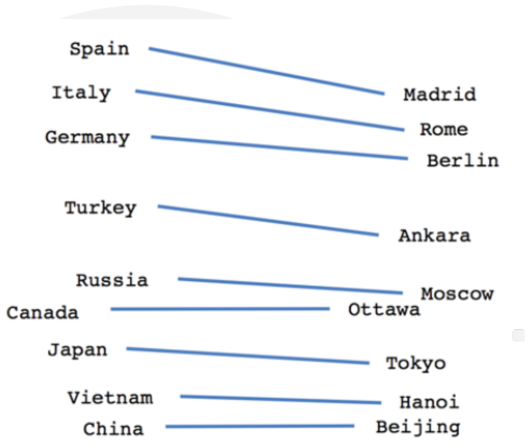
Word2Vec



Male-Female

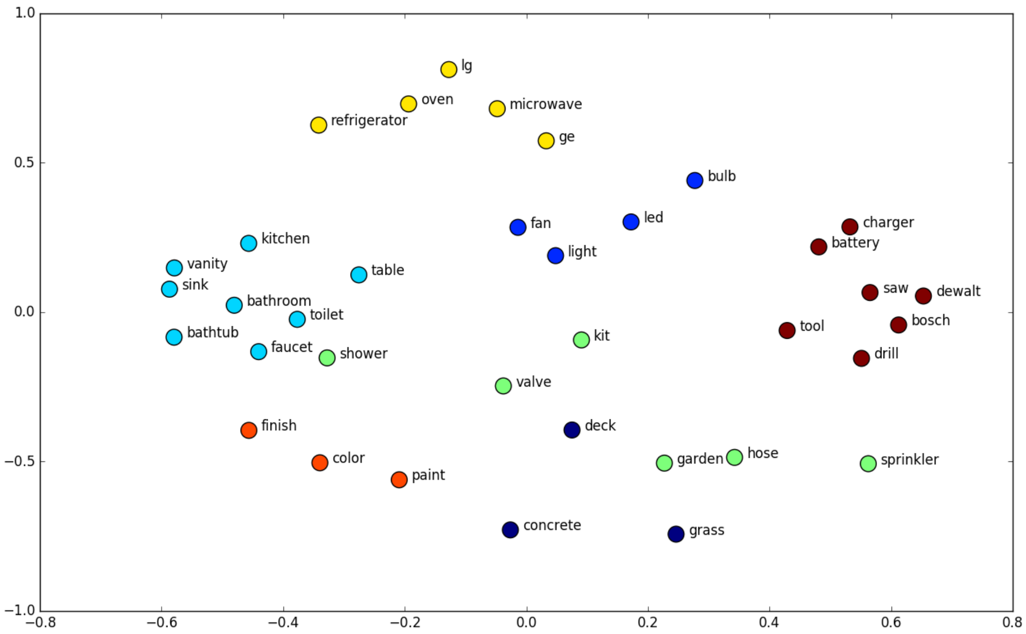


Verb tense



Country-Capital

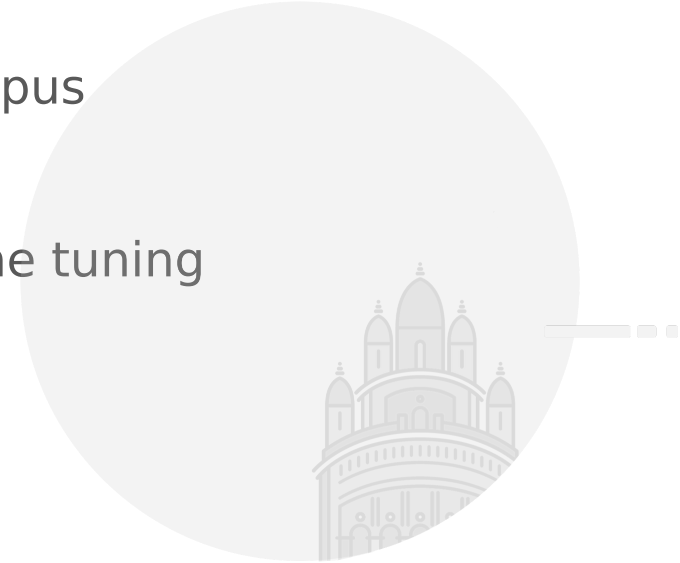
Word2Vec



1.0
0.5
0.0
−0.5
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
lg
oven
microwave
refrigerator
ge
bulb
fan
led
charger
battery
kitchen
light
vanity
table
sink
saw
dewalt
bathroom
toilet
bosch
tool
bathtub
kit
faucet
shower
drill
valve
finish
deck
hose
color
garden
sprinkler
paint
concrete
grass

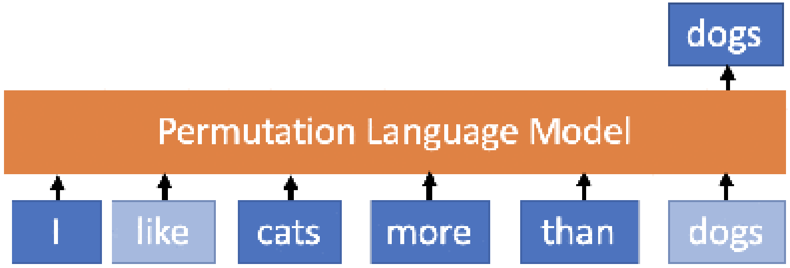
XLNet

- Word2vec
  - Not context sensitive
  - Poor results on retraining on smaller corpus
  -
- Long range dependence
- Pretrained model can be updated via fine tuning
- Bidirectional



XLnet

- Transformer XL
  - Multihead attention layer
    - Takes tensors as input and weights them using attention weights
  - Relative positional embeddings along with word embeddings
  - Permutation in expection to be bidirectional



References

- Multiclass text classification:
  - https://www.kaggle.com/jordiruspira/determining-personality-type-using-ml
  - https://towardsdatascience.com/machine-learning-word-embedding-sentiment-classification-using-keras-b83c28087456
- TF-IDF and LSI:
  - https://nlp.stanford.edu/IR-book/information-retrieval-book.html
  - https://nlp.stanford.edu/IR-book/newslides.html
- Word2Vec:
  - https://en.wikipedia.org/wiki/Distributional_semantics
  - https://web.stanford.edu/class/cs224n/
  - https://www.youtube.com/watch?v=4vT4fzjkGCQ
  - https://www.youtube.com/watch?v=8rXD5-xhemo&list=PLoROMvodv4rOhcuXMZkNm7j3fVwBBY42z&index=2&t=0s
  - https://www.youtube.com/watch?v=kEMJRjEdNzM&list=PLoROMvodv4rOhcuXMZkNm7j3fVwBBY42z&index=3&t=0s
- XLNet:
  - https://github.com/zihangdai/xlnet
  - https://mlexplained.com/2019/06/30/paper-dissected-xlnet-generalized-autoregressive-pretraining-for-language-understanding-explained/
  - https://www.borealisai.com/en/blog/understanding-xlnet/

Thank You

August 4, 2019
Novotel Kolkata Hotel and Residences

GDG Kolkata

WTM : Kolkata :

GDG Cloud Kolkata

GDG DEVFEST KOLKATA