146. List down at least 5 vectorization technique.

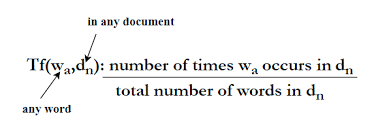
**🡪** 5 vectorization techniques are-

**1. Bag of Words.**

🡪 BOW is a representation of text that describes the occurrence of words within a document. We just keep track of word counts and disregard the grammatical details and the word order. It is called a “bag” of words because any information about the order or structure of words in the document is discarded.

**2. TF -IDF**

🡪 TF stands for Term Frequency.



IDF stands for Inverse Document Frequency. Where DF- Document Frequency.

https://lh4.googleusercontent.com/ACY8Fky4KtK7gZ4RhXvk0W_xdoXB_d-hd-uZuORxoHvF0FZmspY4svXv2IoHYDlJwDBCC97vlt7MiHALNa0hKAg_joZ3K8EkQRcXh_L8cF4zAgFetZZ9StgI8SWi02sIcrV8d_2Q

https://lh5.googleusercontent.com/M5Zkpe89P6lDf4x__CzwQPIMkG7jJ7ediYWVkJYBsv1-eslTAfcCw0zbl5_2xrn3p2sRTgd-budDlwzjgj4lyny-WO5KLGIDosPRRMEj4zR_fNbl5SWEwF0Xm1m8UOK7pTOs6Zxz

**3. Word2Vec**

**🡪** Word2Vec used to group the vectors of similar words together in vector-space.  It uses simple Neural-Networks for word-embeddings.

To implement Word2Vec we use two techniques-

#### I) Skip-Gram , II) CBOW(Continuous Bag of Words.)

### 4. GloVe

### 🡪 It is Similar to Word2Vec but GloVe is also creating contextual word embeddings but given the great performance of Word2Vec.

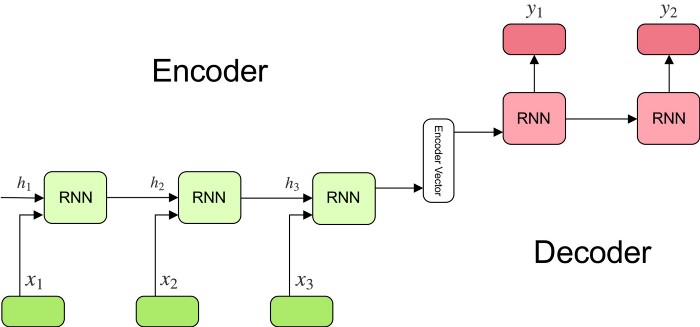
### 5. FastTex

### 🡪 It was introduced by Facebook in 2016. It is also very much similar to Word2Vec. But it has the capability of generalizing the unknown words, which other methods can miss.

147. What is difference between RNN and Encoder-Decoder?

🡪

* A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. It can remembers its input, due to an internal memory.
* Encoder is a stack of several recurrent units where each accepts a single element of the input sequence, collects information for that element and propagates it forward.
* Where Decoder is also a stack of several recurrent units where each predicts an output at a particular time step. each recurrent unit accepts a hidden state from the previous unit and produces and output as well as its own hidden state



148. What do you understand by attention mechanism and what is use of it ?

🡪 In deep learning attention mechanism is an attempt to implement perform a particular action by concentrating a few relevant things, while ignoring the other neural networks. The attention mechanism also brought an improvement over the encoder decoder-based on neural machine translation system in NLP. It helps to  memorize long source sentences in neural machine translation (NMT). Rather than building a single context vector out of the encoder's last hidden state. It was used in other applications, including Computer Vision , speech processing, etc.

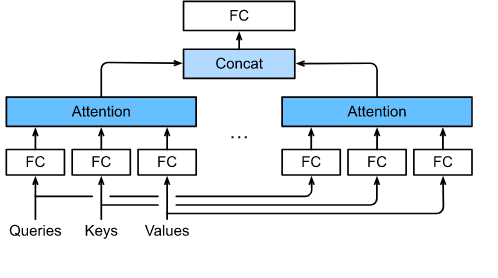
149. Have you read a research paper Attention you all need? If not, then why you are claiming you know NLP?

🡪 I have read it. From that research paper, I understood how attention mechanism is affecting encoder-decoder process in neural machine translation system in NLP.

150. What do you understand by multi headed attention? Explain.

🡪 **Multi headed attention** is a mechanisms which runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension.

Basically, multiple attention heads repeats its computations multiple times in parallel. Each of these is called an Attention Head. The Attention module splits its Query, Key, and Value parameters in multiple ways and passes each split independently through a separate Head. All of these calculations are then combined together to produce a final attention output. It gives the transformer greater power to encode multiple relationships for each word.



1. Tell me something about your project you have done in past?

🡪 In the past I have done several projects from them one of them is, Number plate Detection. In that project, I have used data from Google Open ImageV6 images with OIDV toolkit & trained using YOLOv4-tiny (for real time detection.) & implemented as a local desktop app with flask API & html.

2. What was your Dataset size for ML Project?

🡪 For a ML based project of Flight-Price-Prediction model I have used a dataset with 10684 rows.

3. What is type of your dataset?

🡪 For Flight-Price-Prediction model dataset type was a combination of numerical and string data.

4. What was frequency of your dataset? (E.g. batch, streaming etc)?

🡪 In this dataset, I have used k-fold Cross Validation with 5 folds & used Xgboost algorithm.

5. What was source system for your dataset? (E.g. sensor, satellite Kafka, cloud, etc.).

6. What was kind of derived dataset that you have mentioned in project?

🡪 As a derived dataset, I have used credit card fraud detection data which is a imbalance dataset problem.

7. How you have done validation dataset?

🡪 In Machine learning solutions for robust solution I have used k-fold. If there is n number of folds then, it used (n-1) is for training & the last fold data is for validation.

8. Have you created any pipeline to validate this dataset or you were using any tool?

🡪 For Machine learning solutions I have created pipelines using scikit learn.

9. What do you understand by data lake?

🡪 Data lakes are next-generation data management solutions that can help businesses data scientists in meet big data challenges and drive new levels of real-time analytics. It is an easy accessible, centralized storage repository for large volumes of structured and unstructured for hosting raw, unprocessed enterprise data.

10. What do you understand by data warehousing?

🡪  Data warehouse is a system used for storing and reporting on data. It is basically, data warehousing is an electronic method of organizing, analyzing, and reporting information. In modern business, being able to integrate multiple sources of data is crucial to make better-informed decisions.

A data warehouse essentially combines information from several sources into one comprehensive database.

11. Can you please name some validations that you have done on top of your data?

🡪 K-fold, k-means Clustering, Train-test-split etc.

12. How you have handled streaming dataset?

🡪 I have handled streaming dataset by taking it as a form of batch & storing it in a database.

13. How many different types of environments were available in your project?

🡪 For Number plate detection, I have used multiple environments like- TFOD, OpenCV, Yolo etc. For, different projects, I have used different environments.

14. What was your delivery mechanism for particular project?

🡪 For delivery generally, I use Flask API & OpenCV with HTML & CSS implementation as a local/cloud desktop app.

15. Have you used any OPS pipeline for this current project?

🡪 NO.

16. How you were doing model retraining?

🡪 Using proper pipeline with Py-charm & HTML.

17. How you have implemented model retraining in your project?

🡪 I have used sequence of pipeline so that, whenever it can automatically done all the pre-training processes needed and when start training it will replace the old model with the new one & can do prediction.

18. How frequently you have been doing model retraining and what was the strategy for model retraining?

🡪 For streaming, I am doing model retraining in every 15 days with proper pipeline.