Data Scientist Interview Preparation Topics

Problem Solving – DSA

Get the requirement and the solve, don't directly Jump into Solution optimize the solution and Give O(n) Solution.

1.Given an array of length n, can you find the number of subarrays that sum up to k

2.Find the path from a list of given hops

3.Given the heads of two singly linked-lists headA and headB, return the node at which the two lists intersect. If the two linked lists have no intersection, return None.

4.Given an integer array nums, return an array answer such that answer[i] is equal to the product of all the elements of nums except nums[i]. You must write an algorithm that runs in O(n) time and without using the division operation.

5. 1 1 0 0

0 0 0 1

1 0 1 0

1 1 0 0

Asked to provide the coordinates of the cluster of 1s, given a matrix of 1s and 0s

Given a nested dictionary problem to solve, for finding all the possible key paths.

6. Find the longest non-repeating substring (continuous substring)

string = "aabcbccadef"

output: cadef

7. Print superset of a given set python

S = {1, 2, 3}

8. The cost of a stock on each day is given in an array, find the max profit that you can make by buying and selling on those days.

For example, if the given array is {100, 180, 260, 310, 40, 535, 695}, the maximum profit can be earned by buying on day 0, selling on day 3. Again buy on day 4 and sell on day 6. If the given array of prices is sorted in decreasing order, then profit cannot be earned at All

9. Given a string s, return the longest palindromic substring in s.
10. Given two linked list. Find out if the linked list are intersecting or not

A : a -> b -> c -> d -> ......

B : j -> k -> l -> m -> ......

11. Identify the largest square matrix with all 1s in a given matrix filled with boolean values

[[1., 0., 1., 1., 1., 0., 1., 0., 0., 0.],

[0., 0., 0., 0., 0., 1., 0., 0., 0., 1.],

[0., 0., 0., 0., 0., 1., 0., 1., 1., 0.],

[0., 0., 0., 1., 0., 0., 0., 1., 0., 0.],

[0., 0., 0., 0., 0., 1., 1., 1., 1., 1.],

[1., 0., 0., 0., 0., 1., 1., 1., 1., 1.],

[0., 1., 1., 0., 0., 1., 1., 1., 1., 0.],

[0., 0., 1., 1., 0., 1., 1., 1., 1., 0.],

[0., 1., 0., 0., 0., 0., 0., 0., 0., 0.],

[1., 1., 1., 0., 0., 0., 1., 0., 0., 0.]]

12. Find longest repeating subsequence in a string

13. Nearest neighbour search and Approximate nearest neighbour search.

14. You are given an m \times n matrix where each cell represents the thickness (or weight) of a slab. You need to traverse the matrix from the top-left corner (0, 0) to the bottom-right corner (m-1, n-1). At each step, you can move either:

- Down (i+1, j)
- Right (i, j+1)

The goal is to find a path such that the thickness of the “thinnest slab” along that path is maximized. Return the value of this thickness.

Probability / Statistics

1. Suppose you have 10 pairs of socks, with each pair being a different color. (So, maybe you have a red pair, and a yellow pair, and so forth.) You put them all in the washing machine. The washing machine eats four socks at random. What is the expected value for the number of complete pairs that make it out alive?
2. The cost of a stock on each day is given in an array, find the max profit that you can make by buying and selling on those days.For example, if the given array is {100, 180, 260, 310, 40, 535, 695}, the maximum profit can be earned by buying on day 0, selling on day 3. Again buy on

day 4 and sell on day 6. If the given array of prices is sorted in decreasing order, then profit cannot be earned at all.

3. We have a million text documents which we want to label as "adult"/"non-adult", but we don't have the labels. Still we have to do it. So, as a first step we manually go through 1000 documents and based on what we have read, we find that we can label 970 as "non-adult" and 30 as "adult". So, how do we use this information and use it to label the rest of the documents.
4. You are working for a casino and wish to design simulation for a 2 die game. Write a function to generate the output for the same to be used in simulation. Try and use only one random function.
5. Bias coin P(H) = p. What is the probability of getting atleast 1 H in 20 tosses. What if p -> 0
6. Probability of head for a given coin

Case study:

Intent detection pipeline on tickets

Design Question:

Design a system for tagging photos.Get All Functional, Non Functional requirement and the solve.

ML/DL Topics, Questions

1. how will you evaluate RAG?
2. how do you improve retrieval system ?
3. how do you know which embedding model to choose?
4. Why do we use log-loss?
5. Explain negative sampling in word2vec?
6. If we remove the activation function
7. Vanishing and Exploding Gradients
8. Layer normalization and batch normalization
9. Attention and self-attention
10. Logistic regression is called regression when it is used for classification
11. Difference between softmax and sigmoid?
12. statistical inference, hypothesis testing, metric definition
13. skillset in NLP
14. What are LLMs
15. Transformer Architecture
16. BERT Model
17. Attention Modeling
18. RNN
19. prompting techniques, chunking strategy, usage of different tools used, hallucinations encountered
20. self-attention vs multi-head attention
21. Pre-training tasks used in BERT and Roberta
22. batch normalisation and why is it important

23. residue layer
24. word2vec explanation
25. general encoder decoder process and the parameters that affect the selection and generation of tokens
26. What is loss function for multi-label classification?
27. What's the main difference between a Dense Neural Network and a CNN?
28. Let's assume both dense and cnn have same number of layers? Which will take less space on disk?
29. What’s naive assumption in Naive Bayes ?
30. What are support vectors ?
31. What is LSTM? How does LSTM help us solve vanishing and exploding gradient?
32. What is vanishing gradient and why does it happen?
33. What is p-value? Explain its usage in a simple linear regression problem.
34. PCA, which matrix to use for computing eigenvectors:
35. TSNE
36. detail about how skip-gram word vectors are obtained.
37. What's the difference between a Logistic Regression and an SVM model? When would you prefer one over the other?
38. Bias variance tradeoff, decision tree and random forest
39. Logistic Regression vs SVM: a. Difference: b. When to use:
40. LayoutLM
41. loss function
42. temporal data leakage or any other time series related concept.
43. Train, val, test set necessity
44. difference between L1 and L2?
45. types of loss functions in multiclass classification
46. which to prefer for skewed and multiclass between bert, xgboost
47. how would you train xgboost for text?
48. why cos-sin in positional encodings ?
49. Your model performs really well on the test set but poorly in production.
50. What are your hypotheses about the causes?
51. a. data drift
52. How do you validate whether your hypotheses are correct?
53. overfitting, underfitting, Precision, Recall
54. Activation function, drop out, regularization, attention mechanism, Positional Encoding.
55. How to train a model for OOD detection?
56. How does ReLU help us solve vanishing and exploding gradient?
57. detail about how skip-gram word vectors are obtained. -why sftmax will be computationally intensive and how negative sampling solves for it...
58. Why cos-sin in positional encodings
59. For unit vectors, which of the three similarity metrics would be the best one to use and when? a. Dot Product b. Cosine c. L2
60. Evaluation of RAG
61. GOLDEN SET
62. Use BLEU, ROUGE, METEOR

63. Semantic chunking
64. Why do we use activation in NN?
65. Decoding strategies topk, topp, how Temperature param works
66. Sklearn Random Forest and TF-IDF. Decision Trees, RF classifier and TF-IDF, and approaches to tune the hyperparams.