BarRaiser Interview Experience

➢ Coding (ML)

1. Autoregressive Token Generation

Problem: https://colab.research.google.com/drive/1DEZaH6gFw9b70pweLr_yCENbc9dF-ZHC?usp=sharing" rel="noopener noreferrer" target="_blank">Collab Notebook

We have a model that predicts the next token based on the previous N. Implement a generate method that takes a sequence of tokens and predicts the next N tokens. It should do this auto regressively using the top-k =1.

Solution:
https://colab.research.google.com/drive/1EGM1W4pJ4JCV2Sbs05Ov0DBO6GzvFLfm?usp=sharing" rel="noopener noreferrer" target="_blank">Collab NoteBook

2. Custom Loss Implementation

Problem: https://colab.research.google.com/drive/1CzUJ4KVmbSkRHCs-9lWOW1_ieSeivQpj?usp=sharing" rel="noopener noreferrer" target="_blank">Collab NoteBook

You are working on a computer vision task where you need to implement a custom loss function that combines the Mean Squared Error (MSE) loss with a structural similarity (SSIM) index. The MSE should have a weight of 0.8, and the SSIM should have a weight of 0.2. The SSIM index is calculated using the torchvision library. Implement the custom loss function and demonstrate how it would be used in a training loop with a simple convolutional neural network.

Solution Link: https://colab.research.google.com/drive/1JbEpU2QV7GzXchcNVjwI3DZ-0dlqmnry?usp=sharing

3. Image transformation using OpenCV

Problem: https://colab.research.google.com/drive/19Fb3SY-xsJl6IhlFwtsHO59t2cFckG0t?usp=sharing" rel="noopener noreferrer" target="_blank">Collab Notebook:

Given a kernel, develop a function that applies the kernel to an image. It should be used without padding and with a stride of 1. If input image size is N * M, the output image size should be (N - 2) * (M - 2) with a kernel size 3 * 3

Solution Link:
https://colab.research.google.com/drive/1FTSoCDvAxWsO7xREotZyQjzGlB9E0c5Y?usp=sharing

➢ DS Python Programming

Dataset: Airbnb dataset.ipynb - Colab (google.com)

Solution link: Solutions - Google Docs

Q1: What is the neighborhood in which superhosts have the biggest median price difference with respect to non superhosts? Use the following three columns in the 'listings' dataset to answer this question: 'host_is_superhost', 'neighbourhood_cleansed', and 'price'.

Q2: Which of the review scores has the highest correlation to price? Use the following review score columns in the 'listings' dataset: 'review_scores_rating', 'review_scores_accuracy', 'review_scores_cleanliness', 'review_scores_checkin', 'review_scores_communication', 'review_scores_location', 'review_scores_value'.

Q3: What is the average price difference between a professional host and a non-professional one? Consider a host as professional if they have listings in more than 5 different locations (location is defined by the 'neighbourhood_cleansed' column).

Q4: What is the median price premium given to entire homes / entire apartments with respect to other listings of the same neighborhood? Report the average across all neighborhoods. Use the 'room_type' column in the 'listings' dataset to distinguish between entire homes / entire apartments and other types of listings

Q5: What is the listing with the best expected revenue based on the last 12 months, considering 60% of guests leave reviews and every guest will stay only the minimum number of nights? Use both the 'listings' and 'reviews' datasets for this question and only use listings with minimum nights of stay <= 7. The 'minimum_nights' column indicates the required minimum number of nights of stay for any listing.

Q6: What is the average difference between review scores of superhosts vs normal hosts? Use the 'review_scores_rating' column for determining the average review scores

Q7: Which host attribute has the second-highest correlation with the number of reviews of the listing? Use the following columns as the host attributes: 'host_since', 'host_listings_count', 'host_identity_verified', 'calculated_host_listings_count', 'host_is_superhost', Use 'number_of_reviews' as the column to find correlation with

➢ ML Questions

EASY QUESTION (At least3)

Why is data cleaning necessary before analysis? (DS/DA - Easy)
What does the term "epoch" mean in machine learning?
Explain the difference between precision and recall. In which scenarios would you prioritize one over the other?
How would you split a dataset into training and testing sets using Scikit-learn?

What is the purpose of regularization in machine learning?
What is the difference between supervised and unsupervised learning? Can you provide an example of each?
How do you handle missing values in a dataset?
When performing k means clustering how do you choose K
What is overfitting in machine learning, and how can it be prevented?

Moderate Level (At least 2):
How is K-nearest neighbors (KNN) different from K-means clustering? (DS/DA - Medium)
What are the advantages of using ensemble techniques? Provide examples. (DS/DA - Medium)
What do you understand by Confusion Matrix? (DS/DA - Medium)
What is an F1 score and when would you use it? (DS/DA - Medium)
How might you avoid overfitting in a model? (DS/DA - Medium)
What is the difference between bagging and boosting? (DS/DA - Medium)
What is Maximum Likelihood Estimation (MLE)? (DS/DA - Medium)
What is PCA?
What is backpropagation? Describe a backpropagation workflow in either Pytorch or Tensorflow?

Hard Level:

What is the primary difference between L1 and L2 regularization?

- **Math Questions:**

Question Link: https://t.ly/BE4qI