Data Science Interview Assignment

Submission Deadline

As a guideline, anticipate spending approximately 6 hours on this exercise. The assignment doesn't have to be completed all at once but ideally should be returned within 4 days. Once finished, submit your questions and results to the recruiter you're working with.

Assignment Description

Parspec is revolutionizing the sale of building construction products worldwide, digitizing and organizing the industry's product data, amounting to $5 trillion annually. As part of enhancing product understanding, we seek to determine if a product document pertains to a lighting product or not. Your initial task as a Data Scientist is to construct a model for classifying a product PDF page into either a lighting or non-lighting product.

Data Description:

Training Data

The provided training data is a CSV file named `parspec_train_data.csv`, comprising three columns, serving as your training dataset.

- ID: PDF ID
- URL: The URL where the PDF is hosted; you are required to download the PDFs from this source.
- Is Lighting Product? - Gold Label

Test Data: `parspec_test_data.csv`

The Task:

1. Construct a pipeline to extract text from PDFs.

2. Develop a model for predicting the product type, i.e., whether it is a lighting or non-lighting product.

3. Establish an inference pipeline where any user can input a PDF URL, and the pipeline should return the label (lighting or non-lighting) along with class probabilities. This can be achieved by either creating a small function or developing a hosted pipeline.

4. Make predictions on the test data

Deliverables

The following must be submitted:

1. Code that you wrote to solve the problem
2. Inference pipeline function or hosted app link
3. And, answer to the below questions:
    a. How long did it take to solve the problem?
    b. Explain your solution?
    c. Which model did you use and why?
    d. Any shortcomings and how can we improve the performance?
4. Report your accuracy on test data