Machine Learning Engineer 2 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 either of the 4 category classes.

4 classes names - [Lighting, Fuses, Cables, Others]

Your initial task as an MLE-2 candidate is to construct a model for classifying a product PDF into one of these 4 classes.

Data Description

Please refer to the excel datasheet attached in the folder.

Sheet 1 - train_data : Excel File comprising 2 columns, serving as your training dataset.

- datasheet_link: URL of the hosted PDF

- target_col: The target class of the respective PDF

Sheet 2 - test_data : Excel File comprising 2 columns, serving as your test dataset.

- datasheet_link: URL of the hosted PDF
- target_col: The target class of the respective PDF

The Task

1. Construct a pipeline to extract text from PDFs.

2. Develop a model for predicting the class of product type, i.e. whether it is Lighting, Fuses, Cables, or Others Classes.

3. Establish an inference pipeline where any user can input a PDF URL, and the pipeline should return the label (any of the 4 classes) along with class probabilities.

4. This can be achieved by either creating a small function or developing a hosted pipeline.

5. Make predictions on the test data and report the metrics score

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:
    1. How long did it take to solve the problem?
    2. Explain your solution?
    3. Which model did you use and why?
    4. Any shortcomings and how can we improve the performance?

4. Report the model's performance on the test data using an appropriate metric. Explain why you chose this particular metric.

All the best!

ML AI Team @ Parspec