Program Name:

Document and Materials AI/ML Computer Vision Processor

Short Des :

This Project will use best in class open source technologies to combine Natural Language Processing with BERT for predictions and Tesseract OCR for pulling text from documents hosted within an API and accessed with an Agular web application for data visualization. The end goal being to give Analysts a quick and easy tool to read and process documents or materials such as licenses or passports with only images and no copyable or otherwise extractable text.

Impact Area :

Program , Quality, Data , Accounts and Compliance

Project Owner

Suvarna

Business Value:

Creates an automation process for gathering data on documents that use to only ever be able to be processed by a person. In addition allows for NLP based data organization and evaluation that can provide a detailed analysis of the text once extracted for other forms of data evaluation and extraction.

Deliverables

A Python Flask based API that can take in non text based documents such as PDF license photos or passport photos and correctly extract the text from those images.

A angular application to interact with said API for UI based tools such as the upload feature and visualize back results.

A Flask API for performing initial basic NLP on the processed text such as structure analysis or classification of document types.