Problem Statement: Image Classification from PDF

Link to solution - GitHub Repository

The task requires processing a PDF document that contains scanned images of various types of forms. Each image is accompanied by a label at the top of the page indicating its form type. The objective is to identify and classify each image according to its respective form type based on the information provided in the PDF.

OCR-Assisted Form Type Classification Using Regular Expressions

OCR (Optical Character Recognition):

Use OCR software or libraries (like Tesseract OCR in Python) to extract text from the scanned images in the PDF. OCR will convert the scanned images into machine-readable text, which includes the labels at the top of each form.

Regex (Regular Expressions) for Pattern Matching:

Develop regular expressions to identify and extract the form type from the OCR-extracted text. Define patterns based on the expected format or keywords that indicate the type of form (e.g., specific words or phrases that denote form types like "Form Type", etc.).

PDF to Images Conversion Function:

This function takes a PDF file path as input and converts each page into an image format (e.g., JPEG, PNG). These images are saved in a specified output folder for further processing.

Image Directory Listing Function:

After converting the PDF to images, another function lists all the image filenames within a specified directory. This allows easy access to the processed images for OCR and classification.

OCR and Regex Classification Function:

The core function handles each image individually. It uses OCR to extract text from the image and applies regular expressions to identify and classify the form type based on extracted text patterns.

Print Function for Image and Form Type:

This function prints out the name of each image along with its identified form type. It serves as a verification step to ensure accurate classification of each scanned form.

Segregating Known and Unknown Form Types

To fulfill the task of analyzing only the known form types and segregating them from unknown types, follow these steps:

Unknown images: ['page3.jpg', 'page6.jpg', 'page7.jpg']

Known images: ['page0.jpg', 'page1.jpg', 'page2.jpg', 'page4.jpg', 'page5.jpg', 'page8.jpg', 'page9.jpg']

Image Cropping Based on Form Type for Known Images

To enhance the analysis and preparation of known form images for further processing, we employed manual cropping techniques. The goal was to remove unnecessary parts of the images, focusing only on the relevant sections needed for subsequent tasks. Each form type required specific dimensions, which were manually determined and applied to ensure consistency and accuracy in image preparation.

Example Cropping Specifications:

Plan T4A (P), T3, T4RIF, T4A (OAS)

```
cropped_img = img[700:1400, :]
```

RECEIPT:

```
cropped_img = img[450:1850, :]
```

T5008:

```
cropped_img = img[450:1070, :]
```

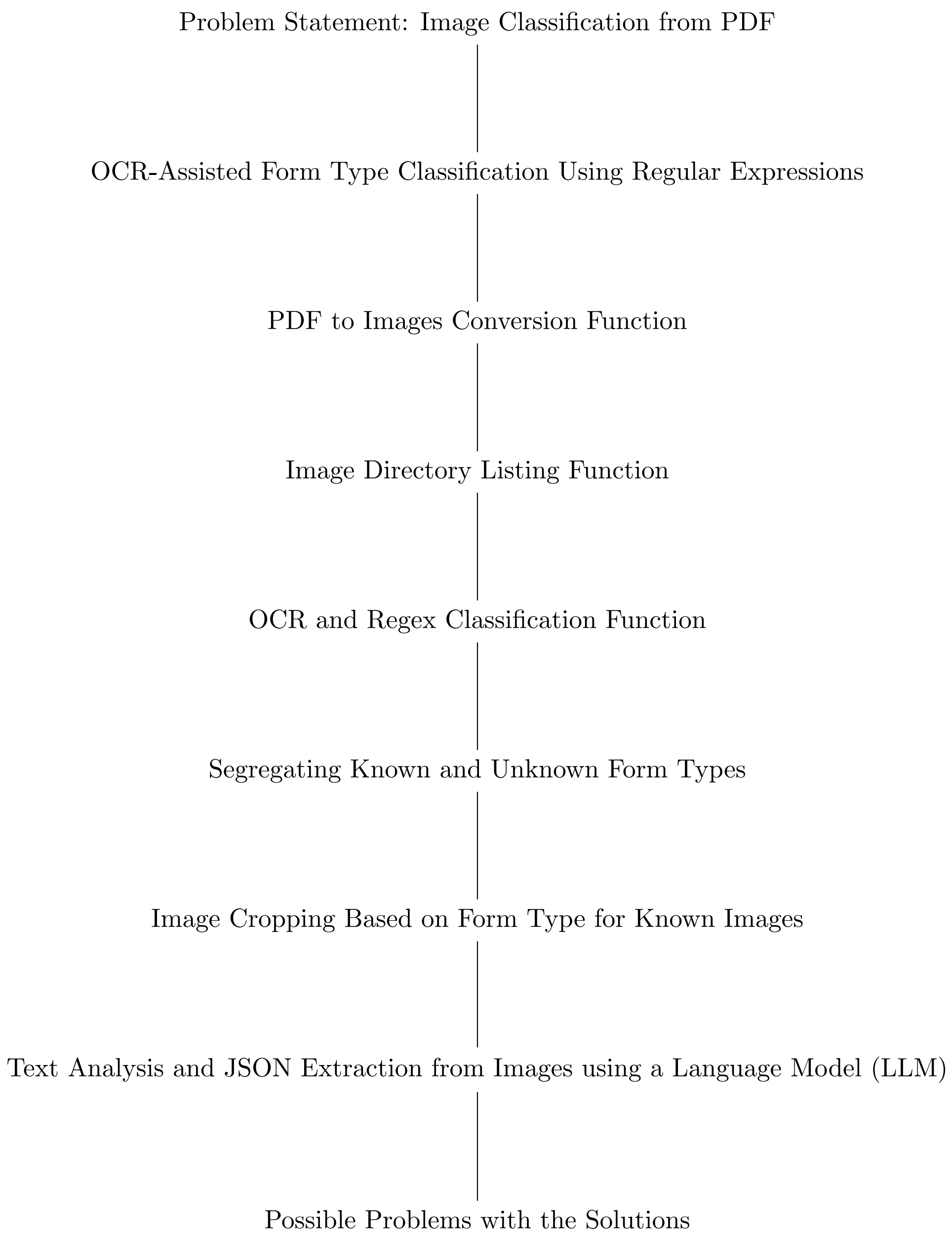
Text Analysis and JSON Extraction from Images using a Language Model

To further streamline the processing of known form images, the next step involves leveraging a Language Model (LLM) for text analysis. The goal is to extract specific information in JSON format from the text content within these images. This approach aims to automate data extraction and organization, enhancing efficiency and accuracy in handling form data.

Possible Problems with the Solutions

Dependency on OCR Accuracy: OCR (Optical Character Recognition) accuracy can vary based on image quality, font styles, and language complexity. New form layouts or variations may require frequent updates to OCR preprocessing and regex patterns.

Prompt Design for LLM: We need to design prompts specific to each form type to guide the model effectively. Each prompt must be unique for every form type. For instance, based on the regular expression, we identify the form type and use the corresponding prompt to extract precise values from the form image.


Problem Statement: Image Classification from PDF
OCR-Assisted Form Type Classification Using Regular Expressions
PDF to Images Conversion Function
Image Directory Listing Function
OCR and Regex Classification Function
Segregating Known and Unknown Form Types
Image Cropping Based on Form Type for Known Images
Text Analysis and JSON Extraction from Images using a Language Model (LLM)
Possible Problems with the Solutions