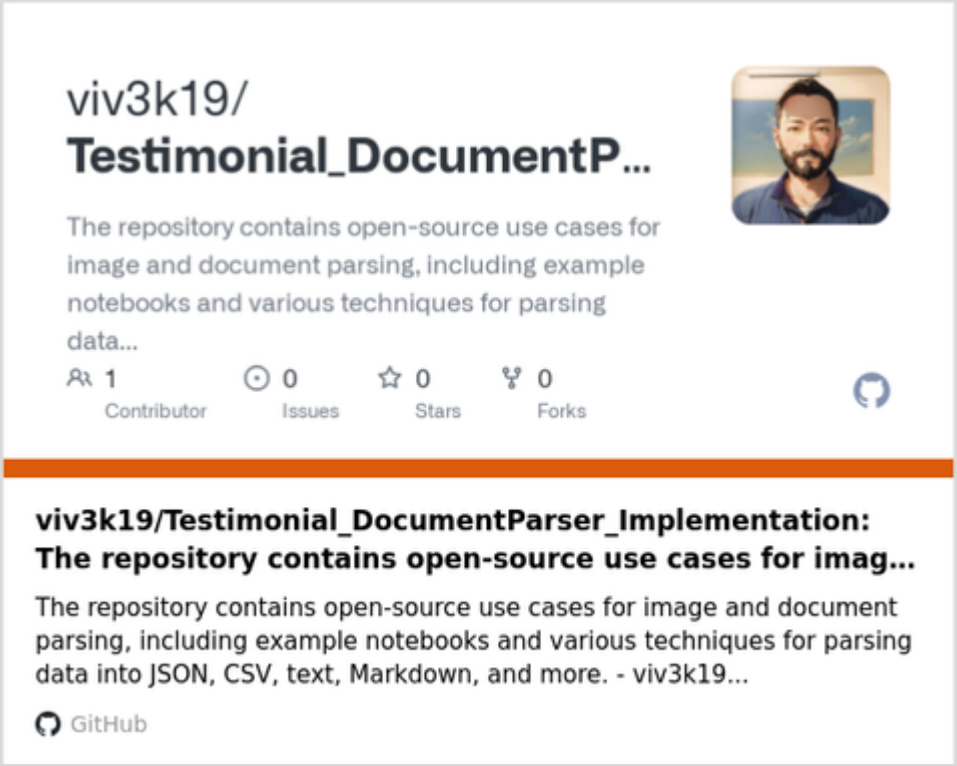
IMAGE
PDF
CV2

JSON
NER
OCR

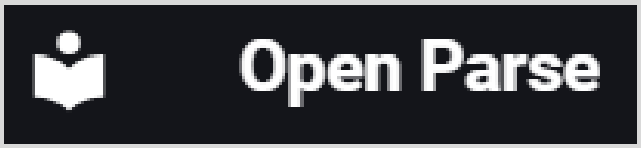
Document Parser Implementation

Project Link



click on the image or here

Research & Experiments



Particular Documentation for each:

click on the logos

How to Use

- Go to my GitHub repository.
- Select the implementation you want.
- Locate the Colab notebook.
- Download the notebook.
- Upload the notebook to Google Colab.
- Run the Colab notebook.
- The dataset is provided in the repository.
- Analyze the Results.

Edgecases

- Varied Layouts
- Noisy Backgrounds
- Mixed Font Styles and Sizes
- Low-Quality Scans
- Handwritten Text
- Complex Tables
- Language and Character Set Variability
- Embedded Graphics and Charts
- Incorrect Document Orientation
- Inconsistent Terminology
- Variable Page Sizes
- Missing Data
- Dynamic Content
- Noise in Input Data
- Overlapping Text

Testcases

- Layout - blockwise extraction
- Metadata extraction
- RegEx based extraction
- OCR based extraction
- API based extraction
- Pattern based extraction
- NER based extraction
- LLM based extraction
- predefined python based extraction

Note: Identified and addressed within notebooks during research and experimentation.

IMAGE
PDF
CV2

JSON
NER
OCR

THANK
YOU