PDF Table Extractor - Project Report

By: Jai Dalmotra
GitHub: jai-dalmotra

1. Introduction

PDF Table Extractor is a Python-based tool that automates the extraction of tables from PDF files into Excel format. It handles both spatial extraction using PyMuPDF and text-based extraction using pdfplumber.

2. Objectives

• Automate PDF table extraction.

• Export extracted data into Excel format.

• Handle both spatial and text-based tables.

• Minimize manual intervention.

• Create a reusable and scalable tool.

3. Technology Stack

• PyMuPDF → Spatial data extraction.

• pdfplumber → Text-based table extraction.

• pandas → Data organization.

• openpyxl → Excel export.

• numpy → Numerical operations.

4. Methodology

1. Read PDF with fitz and pdfplumber.

2. Extract spatial and text-based tables.

3. Export extracted data to Excel.

5. Features

• Automated extraction of multi-page PDF tables.

• Dual extraction methods: spatial and text-based.

• Excel export.

• Row clustering and sanitization.

• Handles multi-line rows.

6. Sample Output

Example Excel Output:

| Name | Age | City |

|----------|------|------------|

| John Doe | 28 | New York |

| Jane Roe | 34 | San Francisco |

7. Challenges and Solutions

Challenges:

• Complex table layouts.

• Inconsistent formatting.

• Multi-line rows.

Solutions:

• Row clustering.

• Character sanitization.

• Dynamic row handling.

8. Benefits

• Automates PDF table extraction.

• Efficient multi-page handling.

• Reduces manual effort.

• Flexible and reusable.

9. Future Enhancements

• Image extraction support.

• Export to CSV and JSON.

• Batch processing with CLI.

• Enhanced OCR with Tesseract.

10. Conclusion

PDF Table Extractor automates table extraction from PDFs into Excel, saving time and effort. It is efficient, reusable, and adaptable for batch processing.

11. References

• PyMuPDF Documentation: https://pymupdf.readthedocs.io

• pdfplumber Documentation: https://github.com/jsvine/pdfplumber

• pandas: https://pandas.pydata.org

• openpyxl: https://openpyxl.readthedocs.io

• NumPy: https://numpy.org