

PDF TABLE EXTRACTION TOOL

By Deepshikha

4 April 2025

INTRODUCTION

- Overview of the project
- Purpose: Extract tabular data from PDFs into structured formats
- Challenges: Handling PDFs with or without borders, irregularly shaped tables

FEATURES

- Extract tables from system-generated PDFs
- Supports PDFs with and without table borders
- Multi-threaded processing for speed
- Outputs data in Excel (.xlsx) and CSV (.csv) formats
- Error handling and logging for robustness

TECH STACK

- Python 3.8+
- Libraries Used:
- pdfplumber: Extract structured text from PDFs
- PyPDF2: Alternative text extraction
- pandas: Data processing and table management
- openpyxl: Excel file handling
- tqdm: Progress tracking
- Optional: Flask/Django for web app version

SYSTEM ARCHITECTURE

- Input: System-generated PDFs (scanned PDFs not supported)
- Processing:
 - Extract words from PDFs using pdfplumber/PyPDF2
 - Group words into table rows
 - Convert grouped data into structured tables
- Output: Tables saved in Excel or CSV format

FOLDER STRUCTURE

- project_folder/
 - | ——— main.py # Main script
 - | ——— extractor.py # Extraction logic
 - | ——— utils.py # Helper functions
 - | ——— requirements.txt # Dependencies
 - | ——— samples/
 - | | ——— input/ # Store PDFs here
 - | | ——— output/ # Extracted tables saved here

INSTALLATION & SETUP

- 1** Prerequisites: Python 3.8+ installed
- 2** Install dependencies:
 - `pip install -r requirements.txt`
- 3** Run the script:
 - `python main.py`
- 4** Results saved in: `samples/output/`

DEMO AND WORKFLOW

1. Upload a PDF to samples/input/
2. Run the script
3. Extracted table data appears in the samples/output/ folder
4. Open Excel/CSV file to verify extracted tables

```
(venv) PS C:\Users\Deepshikha\Desktop\Scoreme_assignment\pdf_table_extractor> python main.py
>>

📄 Processing: test3 (1) (1).pdf
Processing PDFs: 0%|          | 0/1 [00:00<?, ?file/s]
✅ Table saved: samples/output/test3 (1) (1)_page_1.xlsx
✅ Table saved: samples/output/test3 (1) (1)_page_2.xlsx
✅ Table saved: samples/output/test3 (1) (1)_page_3.xlsx
✅ Table saved: samples/output/test3 (1) (1)_page_4.xlsx
✅ Table saved: samples/output/test3 (1) (1)_page_5.xlsx
✅ Table saved: samples/output/test3 (1) (1)_page_6.xlsx
✅ Table saved: samples/output/test3 (1) (1)_page_7.xlsx
✅ Table saved: samples/output/test3 (1) (1)_page_8.xlsx
Processing PDFs: 100%|██████████| 1/1 [00:00<00:00, 6.99file/s]

✅ All PDFs processed successfully!
(venv) PS C:\Users\Deepshikha\Desktop\Scoreme_assignment\pdf_table_extractor>
```


ERROR HANDLING

- Errors logged in: error.log
- Common Issues:
- Missing dependencies → Install with `pip install -r requirements.txt`
- No tables detected → Try a different extraction method
- Output file not found → Check `samples/output/` folder

FUTURE ENHACEMENTS

- ✓ Drag-and-drop GUI for ease of use
- ✓ Web-based version using Flask/Django
- ✓ Enhanced table detection with AI/ML
- ✓ Support for scanned PDFs via OCR integration

The slide features a white background with decorative elements in the corners. The top-left and bottom-left corners have light blue geometric shapes, including rectangles and triangles, some with internal lines. The top-right and bottom-right corners also feature light blue geometric shapes, with the top-right corner including a grid of small circles. The main text is centered in a large, bold, black font.

THANK YOU

Github link

contact: dmmalawliya@gmail.com