PDF to Excel Extraction Tool Documentation

Overview

This tool extracts tabular data from PDFs and saves it as structured Excel files. It scans a given folder for PDFs, processes each file, and generates an Excel file with the extracted table data.

Features

- Processes all PDFs in a specified folder.
- Extracts text while preserving tabular structure.
- Saves extracted data as an Excel file.
- Automatically creates an output folder if it doesn't exist.

Prerequisites

Ensure you have Python installed. The tool requires the following dependencies:

Required Libraries

- `pdfplumber` (for extracting text from PDFs)
- `pandas` (for handling and exporting table data)
- `openpyxl` (for writing Excel files)

Installation

Run the following command to install the necessary packages:

```
pip install pdfplumber pandas openpyxl
```

How to Use

1. Prepare Your PDFs

Place all PDFs you want to process inside a folder named `pdfs` (or any other folder of your choice).

2. Run the Script

Execute the Python script:

python table_extractor.py

3. Output Location

The extracted Excel files will be saved in a folder named `extracted_tables` (created automatically if it doesn't exist). Each Excel file will be named after the corresponding PDF.

Folder Structure

```
project_directory/
│-- table_extractor.py
│-- Input_pdfs/              # Folder containing input PDF files
│-- extracted_tables/     # Folder where extracted Excel files will be saved
```

Example

If `pdfs/` contains:

```
pdfs/
  ├── document1.pdf
  ├── report.pdf
```

After running the script, `extracted_tables/` will contain:

```
extracted_tables/
  ├── document1.xlsx
  ├── report.xlsx
```

Notes

- Ensure PDFs contain structured tabular data; unstructured text may not be properly formatted.
- Adjust the `vertical_threshold` and `gap_threshold` in the script if column alignment needs improvement.

Troubleshooting

1. Missing Dependencies

If you see an error about missing modules, re-run:

pip install pdfplumber pandas openpyxl

2. No Output Generated

- Check if the `pdfs/` folder contains valid PDFs.
- Ensure the PDFs have recognizable text (scanned images may not work without OCR processing).