Document Extraction PDF to CSV OCR Model

The PDF to CSV OCR Text Extraction System is a FastAPI-based web service for uploading and processing PDF files, extracting text, and summarizing key information. It includes database storage for processed file details and generates Excel outputs.

Features

1. PDF to CSV/XLSX conversion
2. Image processing (PDF pages to images)
3. Automatic image rotation
4. Optical Character Recognition (OCR)
5. Supports multiple output formats (CSV/XLSX)
6. Utilizes popular Python libraries

Benefits/Impact

1. Enhanced accessibility of non-editable PDFs
2. Data flexibility in spreadsheet formats
3. Saves time and effort through automation
4. Improved accuracy in text extraction
5. Multi-purpose use of extracted data
6. Customizable and scalable solution