**PDF OCR Text Extraction to Database**

**Table of Contents:**

1. Introduction
2. Project Overview
3. Requirements
4. Installation
5. Configuration
6. Usage
7. Project Structure
8. Dependencies
9. Future Improvements
10. Conclusion
11. **Introduction:**

The PDF OCR Text Extraction to Database Python project is designed to extract text content from PDF documents, perform Optical Character Recognition (OCR) using various libraries such as pdfminer, pytesseract, cv2, PIL, and fitz, and store the extracted information in a MySQL database using the mysql.connector library.

1. **Project Overview:**

The project involves the following key steps:

* Reading PDF documents using pdfminer and PyMuPDF (fitz).
* Preprocessing images with OpenCV (cv2) and PIL.
* Performing OCR on the preprocessed images using pytesseract.
* Storing the extracted text in a MySQL database using mysql.connector.

1. **Requirements:**

Ensure you have the following prerequisites installed:

* Python 3.x
* pdfminer
* pytesseract
* opencv-python
* Pillow (PIL)
* PyMuPDF (fitz)
* mysql-connector-python

1. **Installation:**

* **pip install pdfminer.six**
* **pip install pytesseract**
* **pip install opencv-python**
* **pip install Pillow**
* **pip install PyMuPDF**
* **pip install mysql-connector-python**

1. **Configuration:**

Configure the MySQL database connection in the **mysql connection.py** file.

1. **Usage:**

* Place the PDF files in the **input\_pdf** folder.
* Run the **main.py** script to start the OCR text extraction process.
* Extracted text will be stored in the MySQL database specified in the configuration.

1. **Project Structure:**

**/pdf\_ocr\_extraction**

**│ main.p**

**│ image\_extraction.py**

**│ image\_to\_text function.py**

**│ delete\_images.py**

**│ mysql\_connection.py**

**│ pdf\_function.py**

**│**

**│**

**│**

**├───input\_pdfs**

**│ handwritten.pdf**

**│ java.pdf**

**│ ...**

1. **Dependencies:**

* pdfminer.six
* pytesseract
* opencv-python
* Pillow (PIL)
* PyMuPDF (fitz)
* mysql-connector-python

1. **Implementation Details:**
   1. **PDF Extraction Module:**

**The PDF is opened using PyMuPDF module(fitz) and is extracted into images.**

**Then the image is opened using Pillow Module .**

* 1. **Image Preprocessing**

**PyTesseract is used to set a configuration to read the text in an image.**

* 1. **OCR Processing**

**OCR Processing is done through PyTesseract Module, where the text from the image is extracted and is stored in the database.**

**If the PDF file has no images in it , then pdfminer Module is used to extract the plain text from the PDF and store it in the database.**

* 1. **Database Connectivity**

**Database Connectivity is done through mysql.connector module.**

1. **Future Improvements:**

* Implement multi-threading for faster processing of multiple PDFs.
* Enhance OCR accuracy through better image preprocessing techniques.
* Support for additional databases and storage formats.
* Implement a web-based user interface for better user interaction.

1. **Conclusion:**

The PDF OCR Text Extraction to Database Python project provides a flexible and extensible solution for extracting text from PDFs and storing it in a database. Users can customize the project to meet specific requirements and easily integrate it into their workflows.