Documentation for running the application and about requirements.

# Requirements:

You should have following requirements in your system…

1. Installed python
2. Interpreters (VS code/Pycharm)
3. MySql (prefer open-source cross-platform web server(xampp etc))
4. Installed Java setup(for tabula)

# Creating env and setup:

Create env via ..

1. Virtualenv
2. Python –m venv name\_of\_env

Activation of env in cmd of your current directory where env created

* Name\_of env\Scripts\Activate

Hence you have activated your virtual environment

**Installation of required libraries:**

Now to install all required libraries used in this task (application) you just have to run given requirements.txt

file in your cmd where you have activated your env

* pip install -r requirements.txt

This command will install all required libraries

How to run application:

Once all setup,env and installation of libraries don run given below command in cmd :

* uvicorn main:app --reload

This will run entire application and show a message and a link

with http address like

Link:

INFO: Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit)

And message in terminal

“DEBUG:data\_extraction:

Extracted structured Pdf and scraped Data has been stored in the Database with ID: 2”

click on this link and an html page will open in a browser that has a download button a text box for id to download excel file for any id present in the database so that an excel file containing both pdf extracted structured (table) data and scraped data from website will downloaded in your system that later use for ML model.

So the entire application will run by single command by the use fastapi endpoints where we are fetching the data from database, running the scripts,accessing html page as home page for fastapi link and downloading the excel.

**That’s all to run application rest of all scripts I passed comments so that someone can understand the flow of fastapi ,db integration, exception handling and validations ,logging, pdf structured data extraction using tabula and web scraping scripts by beautifulsoup.**

**Regards:**

**Abdul Manan, ML Engineer**