**DOCUMENTATION**

PROBLEM STATEMENT-3

Data Extraction

This application converts PDF files to CSV format, extracting various types of customer and shop details, including those indicated by tick marks, cross marks, or shaded boxes as per the problem statement in the website itself giving the result of original csv and prompted csv data.

A screenshot of a computer

Description automatically generated

Tech Stack

**1.Python3.8 or higher, pip (installs python packages from terminal)**

**2.Streamlit**

* **Purpose**: Creates the web interface for uploading PDFs, configuring settings, and displaying results.
* **Installation:** pip install streamlit

**3.pytesseract**

* **Purpose**: Performs OCR to extract text from images.
* **Installation**: pip install pytesseract

**4.Pillow (PIL)**

* **Purpose**: Handles image processing.
* **Installation**: pip install Pillow

**5.pdf2image**

* **Purpose**: Converts PDF files to images.
* **Installation**: pip install pdf2image

**6.configparser**

* **Purpose**: Manages configuration settings from config.ini.
* **Installation**: pip install configparser

**7.python-dotenv**

* **Purpose**: Loads environment variables from a .env file.
* **Installation**: pip install python-dotenv

**8.google-generativeai**

* **Purpose**: Integrates with the Gemini API for analyzing CSV content.
* **Installation**: pip install google-generativeai

**External Tools**

**1.Poppler**

* + **Purpose**: Required by pdf2image for PDF to image conversion.
  + **Installation**: https://github.com/oschwartz10612/poppler-windows/releases/

**2.Tesseract-OCR**

* **Purpose**: Performs OCR on images.
* **Installation**: https://sourceforge.net/projects/tesseract-ocr.mirror/

**Set the path of tesseract OCR and poppler in config.ini**

**A screen shot of a computer

Description automatically generated**

**(use your system path)**

**Generate your Gemini API key from** <https://ai.google.dev/gemini-api/docs/api-key>

**Setup the python Gemini API key in .env file**

**A black screen with white text

Description automatically generated**

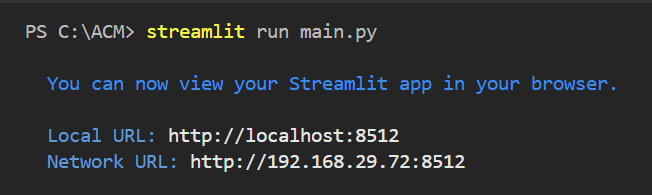
**TO RUN THE CODE:-**

**Put main.py, config.ini and .env file in the same folder**

**Install all of the dependencies**

**In terminal:**

**streamlit run main.py**

****

**##The website will show the details expected as per the problem statement and the csv file converted will directly be downloaded into your system.**