

JA assure PDF EXTRACTOR

Garage GEEKS

Team Members:

Dhruv Bhojraj (RA2211003010522)

Aditya P (RA2211003010548)

Theme: NO. 3 PDF EXTRACTION

August 3, 2024

Contents

1	Project Abstract	2
2	Project Requirements	2
3	Setup Guide	2
4	How To Use The Application	3

1 Project Abstract

The theme of the project is PDF Data Extraction. Since we felt like the theme was vague, we decided to specialize the project to become some sort of program that extracts data from filled forms, and then display statistics about said form on demand. Using Tesseract and CustomTkinter, we finally were able to make the project.

The Github Repository for the project can be found [here](#)

2 Project Requirements

1. To Create a PDF extraction tool that can extract text, ticked boxes and shaded boxes from customer filled pdf forms.
2. To Export the extracted information from the pdf into a .CSV file.

3 Setup Guide

1. Download the directory titled 'pairProgramming' into the local machine.

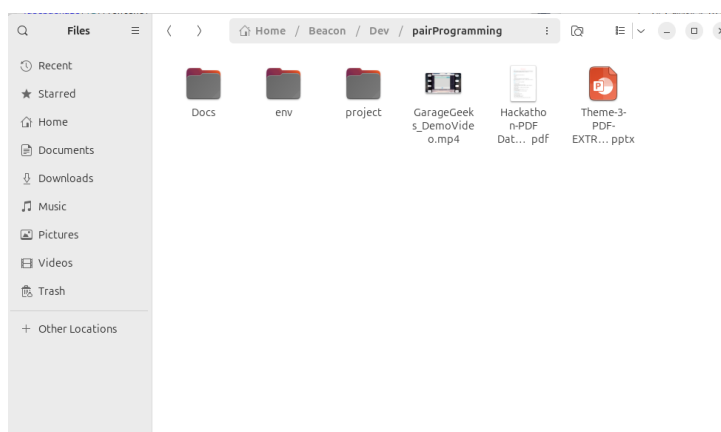


Figure 1: Downloading the project directory

2. Open the existing directory in terminal.

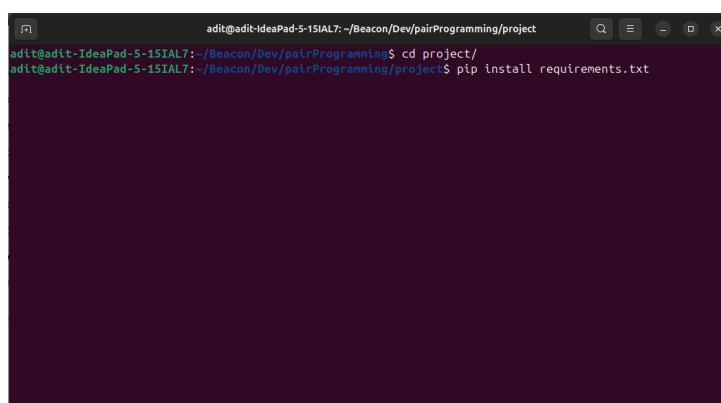


Figure 2: Opening the directory in terminal

3. Type 'cd project' and then 'pip install -r requirements.txt' to install all dependencies required to run the program.
4. Open app.py in an Integrated Development Environment - for example, 'Visual Studio Code' or in terminal: python3 app.py

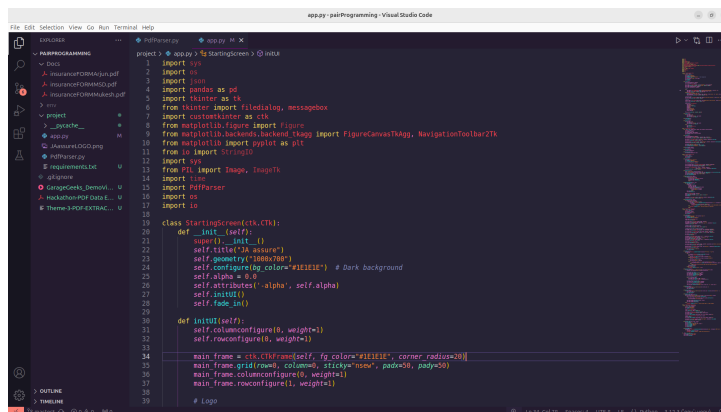


Figure 3: Opening app.py in an IDE

5. Press the Run Button in the top right.



Figure 4: Running the application

4 How To Use The Application

1. Hit 'Get Started Button'.

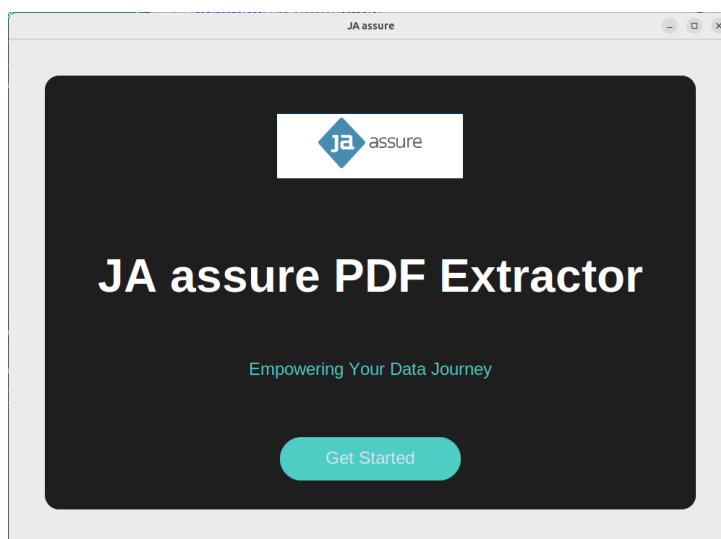


Figure 5: Starting Screen

2. A new window titled 'Profile Selector' appears.

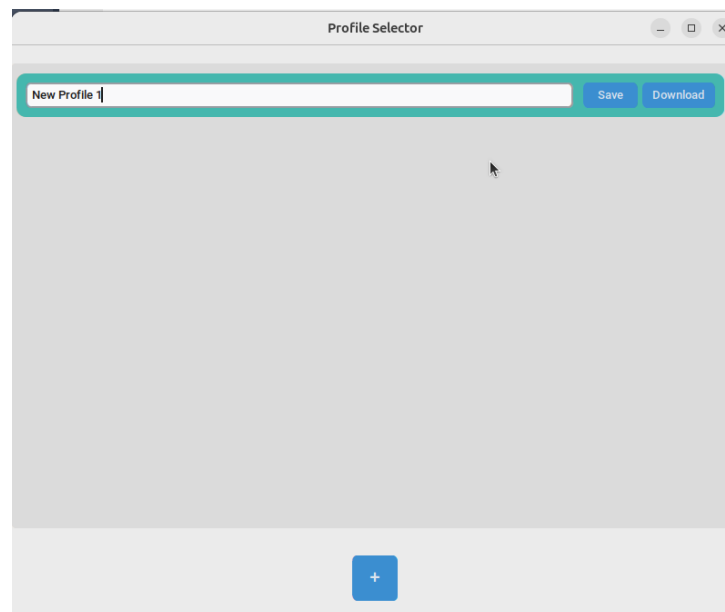


Figure 6: Profile Creation and Selection

3. Press the + button to add any number of profiles you want.

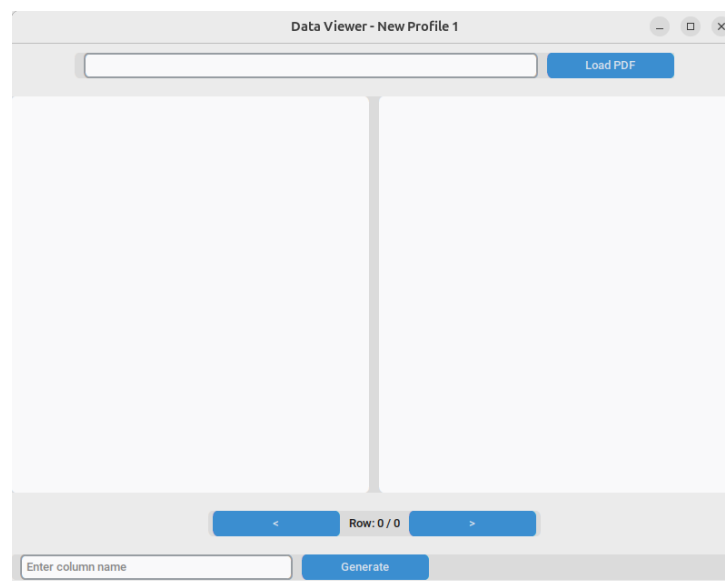


Figure 7: Adding a new profile

4. Type in the Profile Name, like "Medical Malpractice Insurance Applications"

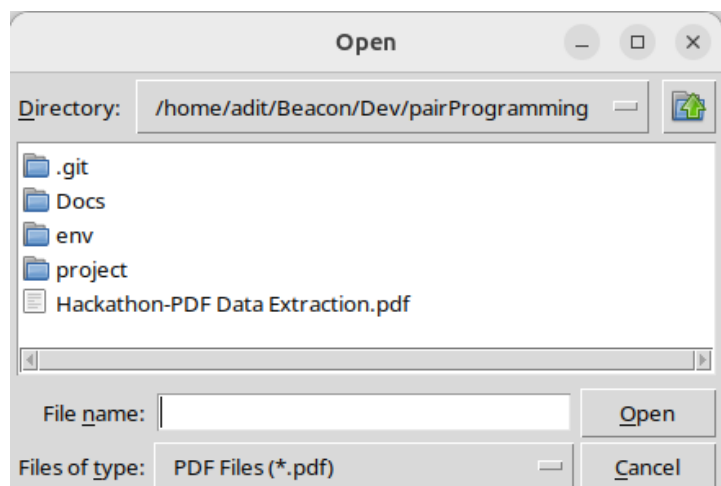


Figure 8: Naming a profile

5. Hit Save.

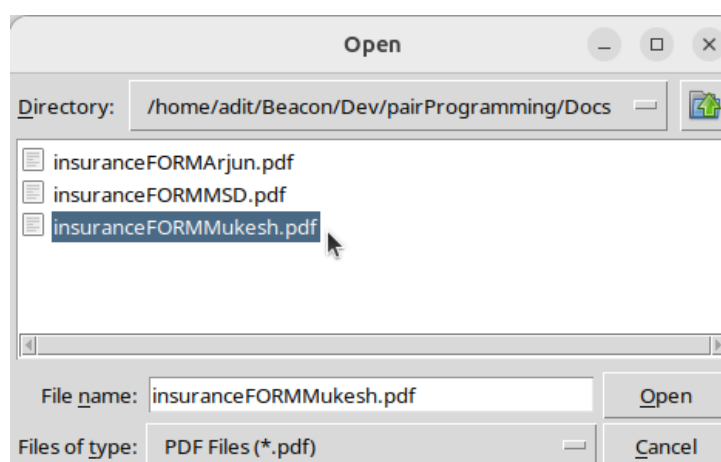


Figure 9: Saving the profile

6. Hit 'edit' to change the name of the profile you have entered.

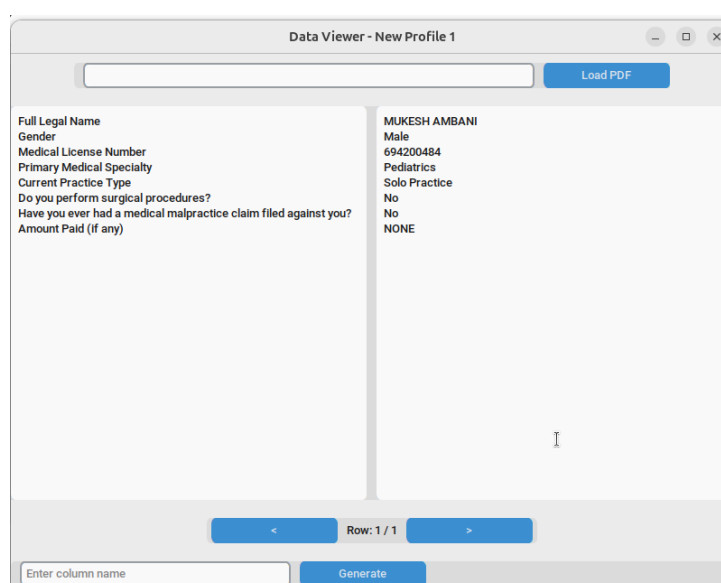
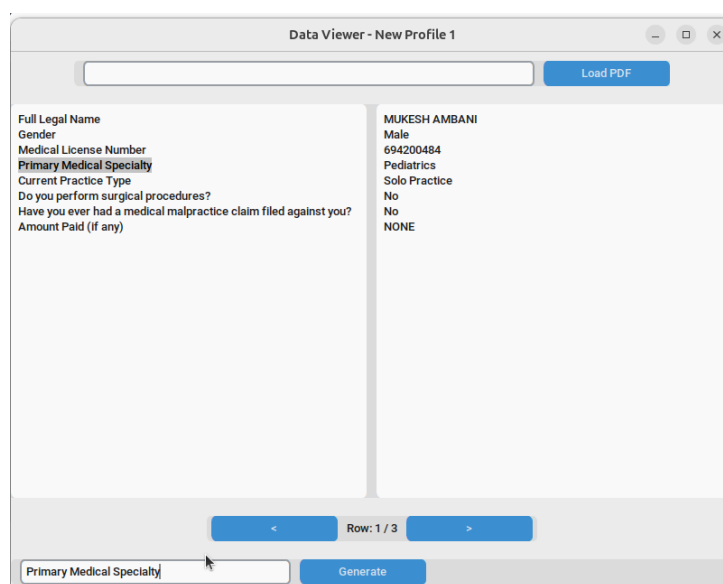


Figure 10: Editing a profile

7. Click on Profile Name to open up a 'DataViewer' window.



Field	Value
Full Legal Name	MUKESH AMBANI
Gender	Male
Medical License Number	694200484
Primary Medical Specialty	Pediatrics
Current Practice Type	Solo Practice
Do you perform surgical procedures?	No
Have you ever had a medical malpractice claim filed against you?	No
Amount Paid (if any)	NONE

Figure 11: Opening the DataViewer

8. Here You can hit 'Load PDF' button to open your file explorer and load up any pdf. Use the PDFs given in the 'Docs' folder to test the model. We have provided a few Documents that emulate medical insurance applications.

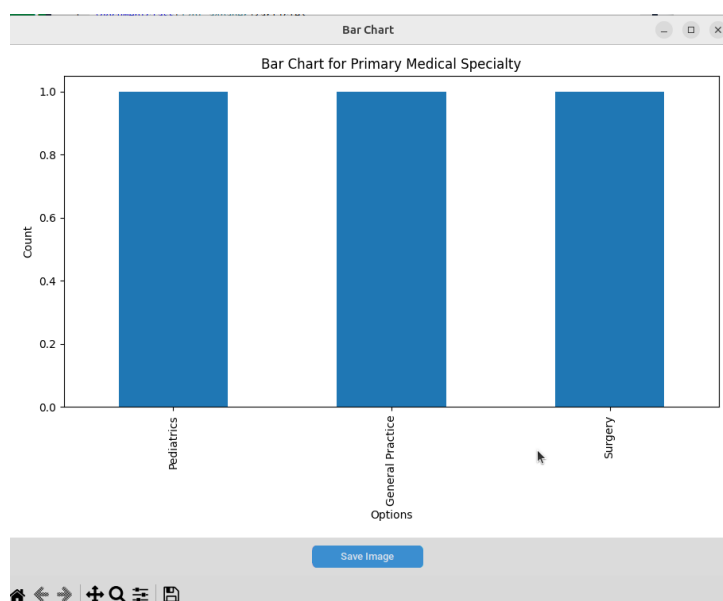


Figure 12: Loading a PDF

9. Data will now be displayed.

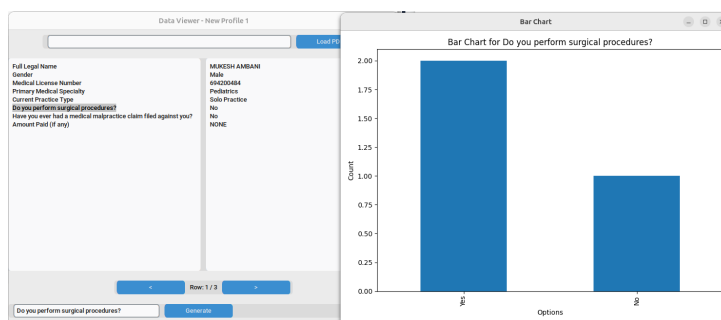


Figure 13: Displaying extracted data

10. Load more PDFs and Use the arrow keys button in the bottom to change the document you are viewing, effectively concatenating the csv.

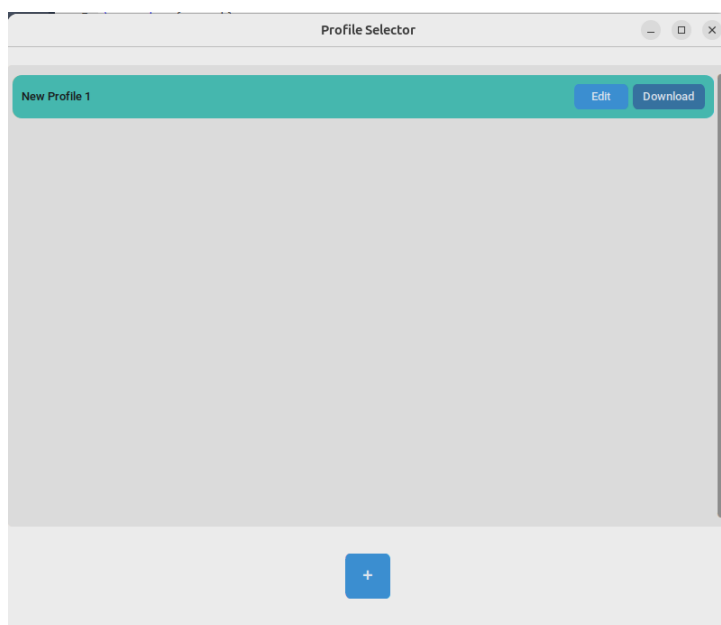


Figure 14: Navigating between documents

11. Type in the name of the field and hit 'Generate' to view the data analytics of the particular field.

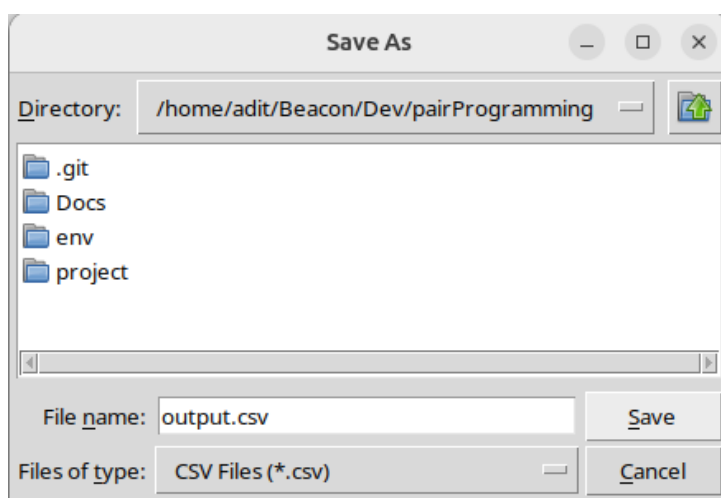


Figure 15: Generating field analytics

12. Press the Download Button in the 'Profile Selector' window to download your CSV file.

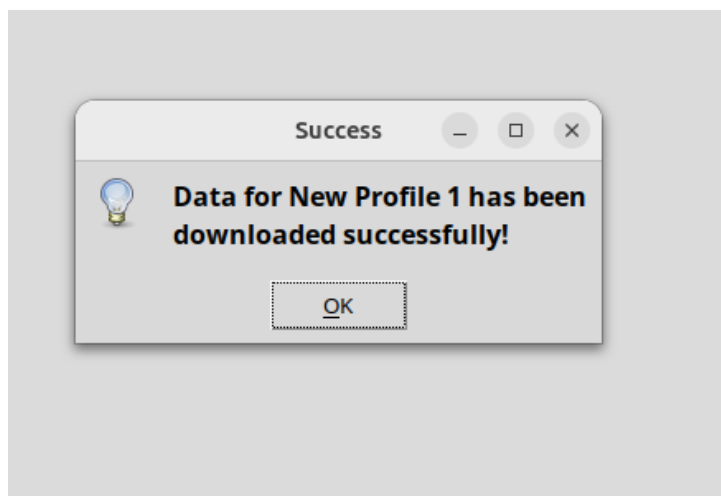
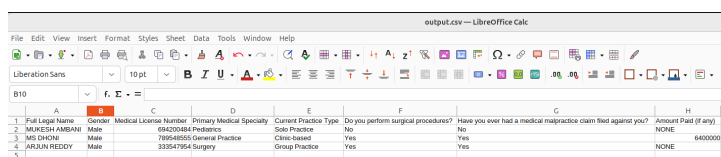


Figure 16: Downloading the CSV file

The CSV in LibreCalc (spreadsheets)



	A	B	C	D	E	F	G	H
1	Full Legal Name	Gender	Medical License Number	Primary Medical Specialty	Current Practice Type	Do you perform surgical procedures?	Have you ever had a medical malpractice claim filed against you?	Amount Paid (if any)
2	MUKESH AMBANI	Male	69450484	Pediatrics	Solo Practice	No	No	NONE
3	MD DINCH	Male	78954855	General Practice	Clinic-based	Yes	Yes	\$400000
4	ARJUN REDDY	Male	33547954	Surgery	Group Practice	Yes	Yes	NONE
5								

Figure 17: Example of exported CSV file in Excel