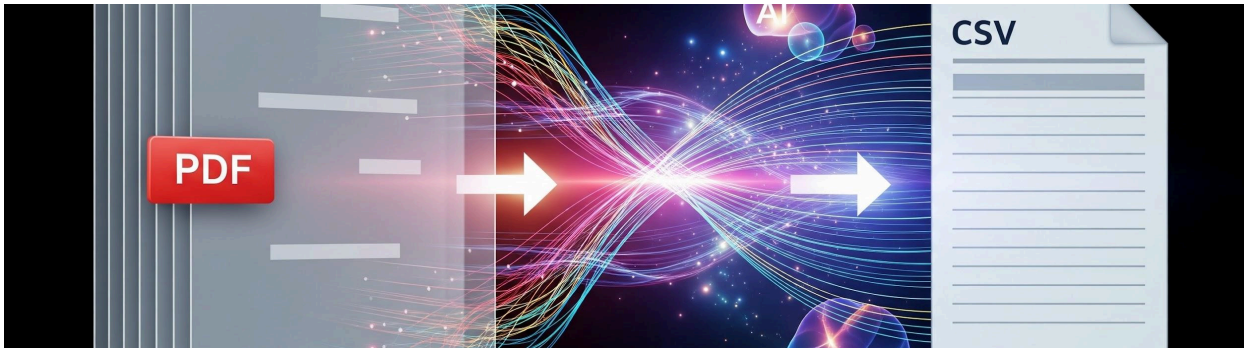

PDF to CSV Program Documentation

This document provides a comprehensive overview of the PDF to CSV program, its functionalities, usage instructions, and post-processing steps for the generated data. This program is designed to extract individual profile data from various company PDFs (e.g., McKinsey, Meta, Goldman Sachs) and convert it into a structured CSV format.

I. Program Overview

This program automates the extraction of individual profiles from PDF documents. It is particularly useful for consolidating information from various company profiles into a single, manageable dataset.

A. Key Features

- **PDF Data Extraction**: Identifies and extracts relevant individual profile information from PDF documents.
- **CSV Conversion**: Transforms the extracted data into a structured CSV file, ready for further processing.

II. Code Explanation

This section details the structure and key components of the PDF to CSV program's codebase.

A. Core Modules

Module Name	Description
`pdf_to_csv.py`	Handles the parsing of PDF documents and the initial extraction of raw data.

B. Data Structures

The program utilizes a dictionary-like structure to temporarily hold extracted data before writing it to the CSV. Each key in the dictionary corresponds to a column in the final CSV, and its value is the extracted data.

III. How to Use the Program

To run the PDF to CSV program, follow these steps:

A. Prerequisites

Before running the program, ensure you have the following installed:

- Python
- Required Python libraries
    - Install via `pip install -r requirements.txt`
    - Or through `pip3 install -r requirements.txt`
- Alter the following sections to the selected company:
    - `PDF_FOLDER`
    - `COMPANY`

  For instance, if the company is McKinsey, change `PDF_FOLDER` to be
  “`mck/`” (if that is the folder name) and `COMPANY` to be “`McKinsey & Company`”

B. Running the Program

1. **Place PDFs**: Ensure all PDF documents containing the individual profiles are located in a designated input directory.
2. **Execute Script**: Run the `pdf_to_csv.py` script from your terminal:

```
Unset
python3 pdf_to_csv.py
```

3. **Output**: The generated CSV file will be saved in the specified output directory: `resume_summary.py`.

IV. Post-Processing the CSV Data

The CSV file generated by this program is considered "unclean" and requires further refinement. The following steps outline the recommended post-processing workflow using ChatGPT to enhance data quality and readability.

A. ChatGPT Integration for Data Cleaning

The following table outlines the key cleaning and refinement tasks to be performed using ChatGPT:

Task	ChatGPT Prompt Guidance	Expected Outcome
Language Translation	"Translate any non-English information in this CSV to English."	All relevant data translated into English.
Data Cleanup	"Clean up the data in this CSV file. Ensure consistency in formatting for names, titles, and company names. Remove any extraneous characters or symbols. There are instances in which the name is inconsistent, so	Consistent and standardized data formatting across all columns.

Task	ChatGPT Prompt Guidance	Expected Outcome
	check the linkedin profile to double check as well."	
Column Refinement	"Refine the following column for better readability: “Work Length”. Make the information easier to understand and digest, keeping a consistent formatting for all"	Improved readability and clarity for designated columns.
Data Validation	"Double-check the information in the csv file to ensure accuracy and correctness based on typical profile data."	Verification and correction of data accuracy in critical columns.

B. Steps for Using ChatGPT

1. **Upload CSV to ChatGPT**: Upload the generated CSV file.
2. **Apply Prompts**: Use the prompt guidance from the table above, adapting them as needed, to instruct ChatGPT on the desired cleaning and refinement.
3. **Review and Save**: Carefully review ChatGPT's output. Once satisfied, copy the cleaned data and paste it back into your CSV file, overwriting the original "unclean" data.

V. Future Enhancements

Potential future enhancements for this program include:

- **Automated ChatGPT Integration**: Developing an API integration with ChatGPT for automated data cleaning.
- **Advanced Error Handling**: Implementing more robust error handling for PDF parsing and data extraction.

- **Efficiency Improvement**: Rewriting the code to be more memory and performance efficient.