**Data verification and Discrepancy Detection**

**Project Summary**

The objective of the project is to develop an auto AI solution that compares the information from the invoices shared by the customer with the data that is present in database (here in excel file) and provide the discrepancies in data.

To provide an end-to-end pipeline that:

* Extracts structured data from invoice PDFs.
* Compares it against validated master data.
* Flags any mismatches (amount or missing items).
* Generates a discrepancy report.
* Optionally, enables this functionality via a **Streamlit web interface**.

**Core Logic & Workflow**

**Input Files**

* Two invoice **PDF files** with itemized data.
* One **Excel file** with reference master data.

**Processing Steps (via**InvoiceComparator**class)**

1. **Extract invoice data** from PDFs using pdfplumber.
2. **Normalize data**: standardize column names, clean whitespace, and parse numbers.
3. **Compare**:
   * Match invoice items to master data using item names.
   * Check for discrepancies in total amounts.
   * Identify missing items.
   * Calculate discrepancy amount and percentage.
4. **Generate report**:
   * Create an Excel file using openpyxl.
   * Apply formatting: bold headers, conditional color highlights, column resizing.

**Why This Approach**

* **pdfplumber** is effective for structured tables in PDF.
* Clean, modular class (InvoiceComparator) design makes the tool maintainable.
* Excel output is user-friendly for finance teams.
* **Streamlit UI** allows non-technical users to interact easily with the tool.

**Streamlit Web Interface**

A minimal UI was created to:

* Upload two PDF files and one Excel file.
* Display the output as a DataFrame.
* Allow downloading the final discrepancy report.

**Running the App via Bash Terminal**

1. Save the app script as app.py.
2. Open your **bash terminal**.
3. Navigate to the script folder:
4. Run the Streamlit app:

streamlit run app.py

1. The browser will open automatically at:

http://localhost:8501

1. Press Ctrl + C in terminal to stop the app.

**Dependencies**

* pdfplumber
* pandas
* openpyxl
* streamlit

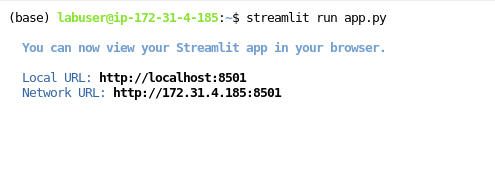
You can install them via pip:

pip install pdfplumber pandas openpyxl streamlit

**Future Improvements**

* Add file validation and error handling.
* Visualize summary statistics on Streamlit.
* Enable multi-invoice comparison with progress feedback.

**Screenshots of Outputs**

****

