Invoice Data Extraction Pipeline

Srivathsan B

20-10-2024

1 Introduction and Motivation

In today's dynamic business landscape, efficiently extracting structured information from invoices is crucial. Invoices often come in various formats, including text-based PDFs, scanned documents, and hybrid PDFs that combine text and images. Each format presents unique challenges for data extraction, impacting financial operations such as cash flow management and expense tracking.

This project aims to develop a cost-effective solution for invoice data extraction while prioritizing accuracy, targeting an extraction rate exceeding 90%. By automating the invoice processing workflow, organizations can significantly reduce manual data entry, allowing them to focus on strategic initiatives.

Leveraging advanced text extraction techniques, we strive to create a scalable system capable of handling diverse invoice formats with minimal human intervention. This solution aims to enhance operational efficiency and improve accuracy in financial management.

2 Source Code

The source code for the Invoice Processing Pipeline is organized in a modular format to enhance readability and maintainability. It can be found in the Github repository- GitHub Repository. Key scripts and modules include:

- `src/pipeline.py`: The main script that executes the extraction process.
- `src/pdf2text.py`: To extract text from PDF.
- `src/text2table.py`: To make structured data out of textual data.
- `src/utils.py`: Contains utility functions for text extraction and data processing.
- `requirements.txt`: A file listing all necessary dependencies for the project.

2.1 Code Documentation

Each function and class in the codebase is documented with:

- **Docstrings**: Explaining the purpose, parameters, and return values of functions and classes.
- **Inline Comments**: Clarifying complex logic and code sections for better understanding.

2.2 Repository Structure

The project directory structure is organized as follows:

```
InvoiceProcessingPipeline/
 src/
    pipeline.py
    pdf2text.py
    text2table.py
    utils.py
 outputs/
    invoice.csv
```

Invoice Number	Invoice Date	Invoice Due Date	Customer Name	Phone	Place of Supply	Address	Taxable Amount	Tax Amount	Total Amount	Total Discount
INV-117	01-Feb-24	29-Jan-24	Naman		23-MADHYA PRADESH		1483.32	183.52	1667	290.02
INV-118	30-Jan-24	30-Jan-24	Rashu		23-MADHYA PRADESH		350	0	350	0
INV-121	29-Jan-24	29-Jan-24	Jitesh Soni	9351333360	27-MAHARASHTRA	Wakad Pune, MAHARASHTRA, 411057	870.93	0	1010	285.94
INV-123	08-Feb-24	08-Feb-24	Asit		23-MADHYA PRADESH		990.46	124.36	1115	152.02
INV-124	10-Feb-24	10-Feb-24	Ankita Sattva		23-MADHYA PRADESH		1125.52	141.32	1115	152.02
INV-127	23-Feb-24	23-Feb-24	Avik Mallick		23-MADHYA PRADESH		943.77	0	944	662.23
INV-128	23-Feb-24	23-Feb-24	Atia Latif		23-MADHYA PRADESH		2076.27	373.72	2450	550
INV-129	23-Feb-24	23-Feb-24	Divya Suhane	6261616609	23-MADHYA PRADESH		1117.05	150.14	1267	172.8
INV-133	01-Mar-24	01-Mar-24	Sheetal Kapur		23-MADHYA PRADESH		2302.15	399.76	2702	471.03
INV-134	01-Mar-24	01-Mar-24	Sheetal Kapur		23-MADHYA PRADESH		723.77	130.28	854	44.95
INV-135	01-Mar-24	01-Mar-24	Mohith Saragur		23-MADHYA PRADESH		691.22	102.22	793	117.28
INV-136	15-Feb-24	04-Mar-24	Rishabh Ramola	7023511429	23-MADHYA PRADESH		961.36	173.04	1134	283.6
INV-138	06-Mar-24	06-Mar-24	Agrani Kandele	8120482988	23-MADHYA PRADESH	vadandana beauty parlour Murawara Katni, MADHYA PRADESH, 483501	1275.34	229.56	1505	308.1
INV-140	06-Mar-24	06-Mar-24	Ankit		23-MADHYA PRADESH		999.36	148.6	1148	286.99
INV-141	06-Mar-24	06-Mar-24	Kasturi Kalwar		23-MADHYA PRADESH		1486.02	267.48	1754	296.5
INV-142	07-Mar-24	07-Mar-24	Urmila Jangam	9552724364	23-MADHYA PRADESH		874.58	157.42	1032	258
INV-143	28-Mar-24	28-Mar-24	Prashant		23-MADHYA PRADESH		6563.98	1048.02	7612	880
INV-144	28-Mar-24	28-Mar-24	Atia Latif		23-MADHYA PRADESH		21914.71	2132.7	24047	933.16
INV-145	28-Mar-24	28-Mar-24	Indraja Mohite		23-MADHYA PRADESH		1917.86	222.94	2141	520.2
INV-146	29-Mar-24	29-Mar-24	Abhikaran Jalonha		23-MADHYA PRADESH		3348.16	528.34	3877	674.48
INV-147	29-Mar-24	29-Mar-24	Divya Suhane	6261616609	23-MADHYA PRADESH		3746.82	268.68	4015	652.31
INV-148	30-Mar-24	01-Apr-24	harshit rathore		23-MADHYA PRADESH		1076.4	157.34	1234	168.24
INV-149	22-Mar-24	01-Apr-24	Karishma Bande	8956804380	23-MADHYA PRADESH		370.64	66.72	437	59.64
INV-150	22-Mar-24	01-Apr-24	Bhusan Naresh		23-MADHYA PRADESH		394.51	71.02	466	63.48

Figure 1: Master Sheet for Invoice Details

```
   purchase_items.csv
Jan to Mar/
   sample invoices/
requirements.txt
README.md
```

3 Technical Details

In this project, we employed `PyMuPDF` for extracting text from PDF invoices. This library provides robust capabilities for handling various PDF formats, allowing us to process both text-based and scanned documents efficiently.

Once the text is extracted, we utilize regular expressions (regex) in conjunction with the known structure of the invoices to carefully and precisely extract essential details. The key data points extracted include:

After extracting the necessary details, we create two CSV files:

- One CSV (invoice.csv) as depicted in Fig.1, file contains customer details along with overall invoice information.
- The second CSV (*purchase_items.csv*) as depicted in Fig.2, file lists items purchased for each specific invoice number.

These two tables are interconnected, with the **invoice number** serving as the **primary key**, ensuring a reliable link between customer details and the purchased items for each invoice.

4 Justification for Chosen Methods

The selection of `PyMuPDF` as the primary tool for text extraction from invoices is grounded in its reliability and efficiency. Unlike other methods that may introduce stochastic behavior, `PyMuPDF` offers a deterministic approach when dealing with well-structured PDF documents. Given that all invoices processed in this project adhere to a consistent structure, this library can handle the extraction task with precision.

By leveraging the known layout of the invoices, we can extract relevant data points accurately, resulting in a trust determination rate exceeding 99%. This high level of trust is crucial for ensuring the integrity of the extracted information. The effectiveness of our approach can be verified by examining

Invoice Number	Item Number	Item Name	Discounted Price	Actual Price	Discount Percentage	Quantity	Taxable Value	Tax Amount	Tax Percentage	Amount
INV-117	1	Kera M 5% Solution	492.86	616.07	-20	1	492.86	59.14	12	552
INV-117	2	Arachitol Nano (60k) 4*5ml	299.58	340.43	-12	3	898.73	107.85	12	1006.58
INV-117	3	Neurobion Forte - 30 tablets	30.58	34.75	-12	3	91.73	16.51	18	108.24
INV-118	1	Vitamin b12 test	350	350	0	1	350	0	0	350
INV-121	1	Acne-UV Gel - spf 50 (50 gm)	581.57	775.42	-25	1	581.57	104.68	18	686.25
INV-121	2	Arachitol Nano (60k) 4*5ml	289.36	340.43	-15	1	289.36	34.72	12	324.09
INV-123	1	Arachitol Nano (60k) 4*5ml	299.58	340.43	-12	3	898.73	107.85	12	1006.58
INV-123	2	Neurobion Forte - 30 tablets	30.58	34.75	-12	3	91.73	16.51	18	108.24
INV-124	1	Arachitol Nano (60k) 4*5ml	340.43	340.43	0	3	1021.29	122.55	12	1143.84
INV-124	2	Neurobion Forte - 30 tablets	34.75	34.75	0	3	104.24	18.76	18	123
INV-127	1	Vitamin b12 test	397.5	750	-47	1	397.5	0	0	397.5
INV-127	2	HBA1C Test	349.27	659	-47	1	349.27	0	0	349.27
INV-127	3	Last bill pending	197	197	0	1	197	0	0	197
INV-129	1	sotret nf 16 mg - 10 capsules	282.86	321.43	-12	3	848.57	101.83	12	950.4
INV-129	2	Ekran Aqua Sunscreen Spf 30	268.47	305.08	-12	1	268.47	48.33	18	316.8
INV-133	1	Glogeous facewash	785.17	923.73	-15	1	785.17	141.33	18	926.5
INV-133	2	Bristaa Intense Cream	749.95	872.03	-14	1	749.95	134.99	18	884.94

Figure 2: Preview of Items Sheet for Purchased items mapped with Invoice Number

the accuracy scores in the generated CSV files. The results demonstrate strong performance across the majority of the invoices provided, validating the robustness of our methodology.

5 Accuracy and Trust Assessment

The effectiveness of this approach can be verified by examining the accuracy scores in the generated CSV files. The results demonstrate strong performance across the majority of the invoices provided, validating the robustness of our methodology.

However, it is important to note that due to the absence of ground truth data structured, we are unable to quantify the accuracy statistically. Nonetheless, through manual inspection, the system appears to perform quite well, consistently extracting key details from the invoices with a high degree of fidelity. This qualitative assessment further supports our confidence in the accuracy and reliability of the extracted information.

6 Performance Analysis

The performance analysis of the invoice processing system reveals significant advantages stemming from our deterministic approach. This method demonstrates high processing speed and low resource utilization, allowing for efficient handling of multiple invoices concurrently. The structured nature of the invoices enables easy parallelization of the extraction process, which further enhances performance and reduces processing time.

An alternative approach considered was the utilization of large language models (LLMs) for data extraction. While LLMs can provide robust capabilities in understanding and processing varied textual data, they tend to be slow and computationally expensive. The overhead associated with using LLMs could introduce delays in processing time, making them less suitable for scenarios where speed and resource efficiency are critical.

7 Future Work

- **Usage of Large Language Models:** As a next step, it will be essential to incorporate LLMs for processing scanned images and JPEG or PNG files, where the structure may not be as clearly defined. However, the known structure of the provided PDF invoices motivated our choice for a

more simplistic and efficient approach in this initial phase. This methodology effectively balances performance, cost-effectiveness, and accuracy in extracting crucial invoice data.

- **PyTessaract Module:** To further enhance the invoice processing system, several improvements can be made. One key area for future work is the incorporation of computer vision techniques to handle image-based invoices more effectively. By utilizing `PyTesseract`, we can improve the quality of text extraction from image files, such as scanned documents or low-resolution PDFs. Preprocessing these images with advanced computer vision techniques will ensure higher-quality text input for further analysis.
- Once the text has been reliably extracted, generative AI models can be employed to extract structured information. This hybrid approach—feeding high-quality image-extracted text into Generative AI models—can improve the overall accuracy of the framework compared to using a plain Large Language Model (LLM) for text-based extraction alone. By combining both computer vision and AI-based text processing, the system can handle a wider range of invoice formats, increasing robustness and accuracy across different document types.

8 Acknowledgments

I would like to express my sincere gratitude to **Zolvit** for providing the problem statement and the relevant data for this project. Their support and resources were instrumental in developing and testing the invoice processing pipeline.