ZOLVIT-PS

Name: Arun Palaniappan
Roll No: me20b036

Introduction

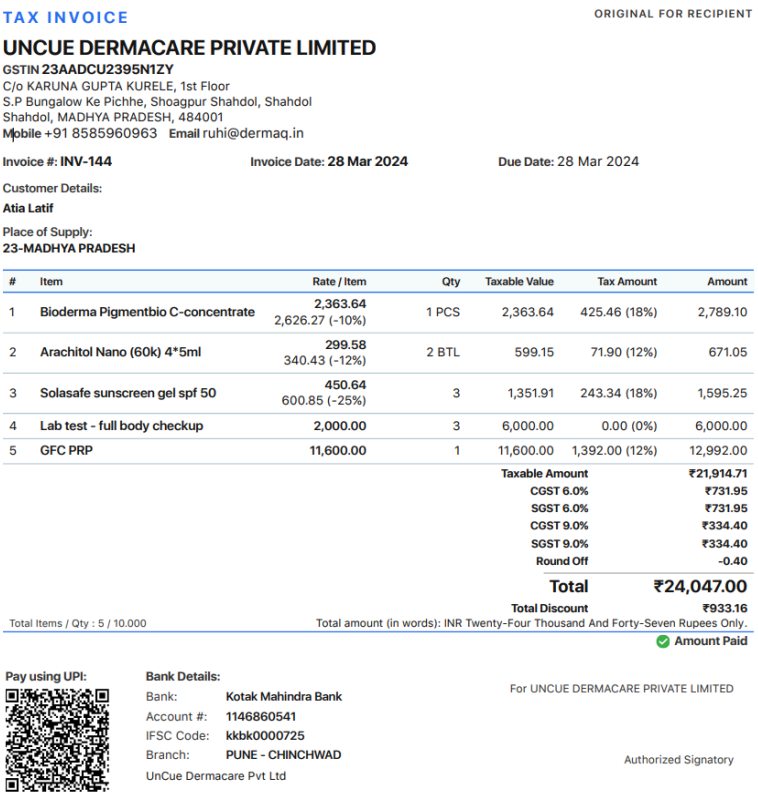
The purpose of this report is to present a comprehensive overview of the Invoice Data Extraction System developed for efficiently extracting and validating information from various types of invoice PDFs. The report covers requirements, deliverables, performance metrics, and an analysis of the implemented solution.

PDF Types

I have used `pymupdf` library to extract the text from the PDF, and it can scan all types of PDFs like regular PDFs (text-based), scanned PDFs (image-based), and mixed PDFs (containing both text and images).

Extracted Data

I have split the data extracted into four broad subcategories, and the data extracted under each sub-category is stored as a separate sheet in the same Excel file for easy access.
'Medicine Bill' - Gives the medical data.
'Personal Information' - Gives personal information like contact details and bank account details.
'Tax and Total Amount' - Gives the total amount and tax details.
'Data Validity' - Gives the validity check of the extracted data.



TAX INVOICE

ORIGINAL FOR RECIPIENT

UNCUE DERMACARE PRIVATE LIMITED
GSTIN 23AADCU2395N1ZY
C/o KARUNA GUPTA KURELE, 1st Floor
S.P Bungalow Ke Pichhe, Shoagpur Shahdol, Shahdol
Shahdol, MADHYA PRADESH, 484001
Mobile +91 8585960963 Email ruhi@dermaq.in

Invoice #: **INV-144** Invoice Date: **28 Mar 2024** Due Date: 28 Mar 2024

Customer Details:
Atia Latif

Place of Supply:
23-MADHYA PRADESH

#	Item	Rate / Item	Qty	Taxable Value	Tax Amount	Amount
1	Bioderma Pigmentbio C-concentrate	2,363.64 2,626.27 (-10%)	1 PCS	2,363.64	425.46 (18%)	2,789.10
2	Arachitol Nano (60k) 4*5ml	299.58 340.43 (-12%)	2 BTL	599.15	71.90 (12%)	671.05
3	Solasafe sunscreen gel spf 50	450.64 600.85 (-25%)	3	1,351.91	243.34 (18%)	1,595.25
4	Lab test - full body checkup	2,000.00	3	6,000.00	0.00 (0%)	6,000.00
5	GFC PRP	11,600.00	1	11,600.00	1,392.00 (12%)	12,992.00

Taxable Amount	₹21,914.71
CGST 6.0%	₹731.95
SGST 6.0%	₹731.95
CGST 9.0%	₹334.40
SGST 9.0%	₹334.40
Round Off	-0.40
Total	**₹24,047.00**
Total Discount	₹933.16

Total Items / Qty : 5 / 10.000

Total amount (in words): INR Twenty-Four Thousand And Forty-Seven Rupees Only.

Amount Paid

Pay using UPI:

Bank Details:
Bank: Kotak Mahindra Bank
Account #: 1146860541
IFSC Code: kkbk0000725
Branch: PUNE - CHINCHWAD
UnCue Dermacare Pvt Ltd

For UNCUE DERMACARE PRIVATE LIMITED

Authorized Signatory

The above is an example PDF(“INV-144_Atia Latif.pdf”) from the given data.
Below, I have attached snapshots of different sheets of the same Excel file containing the extracted data.

Sheet 1: Medicine Bill

Item_list	Rate_Item_list	Qty_list	Taxable_Value_list	Tax_Amount_list	Amount_list
Bioderma Pigmentbio C-concentrate	2,363.64 = 2,626.27 (-10%)	1 PCS	2,363.64	425.46 (18%)	2789.1
Arachitol Nano (60k) 4*5ml	299.58 = 340.43 (-12%)	2 BTL	599.15	71.90 (12%)	671.05
Solasafe sunscreen gel spf 50	450.64 = 600.85 (-25%)	3	1,351.91	243.34 (18%)	1595.25
Lab test - full body checkup	2,000.00	3	6,000.00	0.00 (0%)	6000
GFC PRP	11,600.00	1	11,600.00	1,392.00 (12%)	12992

Medicine Bill | Personal Information | Tax and Total Amount | Data Validity

Sheet 2: Personal Information

Entity	Data
GSTIN	23AADCU2395N1ZY
Address	C/o KARUNA GUPTA KURELE, 1st Floor, S.P Bungalow Ke Pichhe, Shoagpur Shahdol, Shahdol, Shahdol, MADHYA PRADESH, 484001
Mobile	+91 8585960963
Email	ruhi@dermaq.in
Invoice #	INV-144
Invoice Date	28 Mar 2024
Due Date	28 Mar 2024
Customer Details	Atia Latif
Place of Supply	23-MADHYA PRADESH
Total Items	5
Qty	10.000
Total amount (in words)	INR Twenty-Four Thousand And Forty-Seven Rupees Only.
Bank	Kotak Mahindra Bank
Account #	1146860541
IFSC Code	kkbk0000725
Branch	PUNE - CHINCHWAD

Medicine Bill | Personal Information | Tax and Total Amount | Data Validity

Sheet 3: Tax and Total Amount

	A	B	C	D	E	F	G	H	I
1	**Entity**	**Value**							
2	Taxable Amount	21914.71							
3	CGST 6.0%	731.95							
4	SGST 6.0%	731.95							
5	CGST 9.0%	334.4							
6	SGST 9.0%	334.4							
7	Round Off	0.4							
8	Total	24047							
9	Total Discount	933.16							
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									
24									
25									

Medicine Bill | Personal Information | **Tax and Total Amount** | Data Validity

Sheet 4: Data Validity

	A	B	C	D	E	F	G	H	I	J	K	L
1	**mobile**	**email**	**invoice_number**	**invoice_date**	**due_date**	**total_items**	**qty**	**account_number**	**sum_amount_table_round_off**	**sum_amount_tax**	**total_unique_quantity**	**total_quantity**
2	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
3												
4												
5												
6												
7												
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												
25												

Medicine Bill | Personal Information | Tax and Total Amount | **Data Validity**

Accuracy Analysis

I have achieved an overall accuracy of **99.3056%**, as calculated(In the code file) below:

```
Overall_accuracy_score:  99.30555555555556
```

These are the accuracy percentages of each entity:

```
Accuracy Percentages:
mobile: 100.0%
email: 100.0%
invoice_number: 100.0%
invoice_date: 100.0%
due_date: 100.0%
total_items: 100.0%
qty: 100.0%
account_number: 100.0%
sum_amount_table_round_off: 100.0%
sum_amount_tax: 91.67%
total_unique_quantity: 100.0%
total_quantity: 100.0%
```

Scalability and Efficiency

The time taken for the code to process 24 PDFs is **1.419003 seconds** as given below:

```
Time taken: 1.419003 seconds
```

This implies that the Time Taken for 1 PDF is (1.41903)/(24) seconds,
Which is **0.059126 seconds/PDF**

So, if we have 1000 PDFs like this, the time taken would be 59.12650 seconds, which is less than 1 minute for 1000 PDFs.

Logic for Data Extraction

The use of colon ":"

The colon can be effectively used to collect personal information like bank details, email, and mobile number. It is implemented as below:

```
for i, line in enumerate(lines):

    if (":" in line) and (line.strip()[-1] == ":") and (lines[i + 1].strip()[-1] != ":"):
        info[line.strip()[:-1]] = lines[i + 1].strip()

    if (":" in line) and (line.strip()[-1] != ":") and (len(line.strip().split('/'))>=3):
        info[line.strip().split(':')[0].split('/')[0].strip()] = line.split(":")[-1].split('/')[0].strip()
        info[line.strip().split(':')[0].split('/')[1].strip()] = line.split(":")[-1].split('/')[1].strip()

    if (":" in line) and (line.strip()[-1] != ":") and (len(line.strip().split('/'))<3):
        info[line.strip().split(':')[0]] = line.split(":")[-1].strip()
```

Logic for Data Validation

Mobile Number:
We check if there exists 10 digits after “+91” or “country code”.

Email:
We check if the email has “@” and “.” in it.

Invoice Number:
We check if it starts with “INV-”

Total Amount from the Medical Bill:

	A	B	C	D	E	F
1	Item_list	Rate_Item_list	Qty_list	Taxable_Value_list	Tax_Amount_list	Amount_list
2	Bioderma Pigmentbio C-concentrate	2,363.64 = 2,626.27 (-10%)	1 PCS	2,363.64	425.46 (18%)	2789.1
3	Arachitol Nano (60k) 4*5ml	299.58 = 340.43 (-12%)	2 BTL	599.15	71.90 (12%)	671.05
4	Solasafe sunscreen gel spf 50	450.64 = 600.85 (-25%)	3	1,351.91	243.34 (18%)	1595.25
5	Lab test - full body checkup	2,000.00	3	6,000.00	0.00 (0%)	6000
6	GFC PRP	11,600.00	1	11,600.00	1,392.00 (12%)	12992

We take the sum of the rightmost amount column and check if it matches the total amount extracted from the PDF.

Total Amount from Tax table:

	A	B
1	Entity	Value
2	Taxable Amount	21914.71
3	CGST 6.0%	731.95
4	SGST 6.0%	731.95
5	CGST 9.0%	334.4
6	SGST 9.0%	334.4
7	Round Off	0.4
8	Total	24047
9	Total Discount	933.16

We check if the (Taxable amount + Taxes - Discount - Round Off) is equal to the (Total) extracted.

Total Quantity of Items and Total Number of Unique Items:

	A	B	C	D	E	F
1	Item_list	Rate_Item_list	Qty_list	Taxable_Value_list	Tax_Amount_list	Amount_list
2	Bioderma Pigmentbio C-concentrate	2,363.64 = 2,626.27 (-10%)	1 PCS	2,363.64	425.46 (18%)	2789.1
3	Arachitol Nano (60k) 4*5ml	299.58 = 340.43 (-12%)	2 BTL	599.15	71.90 (12%)	671.05
4	Solasafe sunscreen gel spf 50	450.64 = 600.85 (-25%)	3	1,351.91	243.34 (18%)	1595.25
5	Lab test - full body checkup	2,000.00	3	6,000.00	0.00 (0%)	6000
6	GFC PRP	11,600.00	1	11,600.00	1,392.00 (12%)	12992

11	Total Items	5
12	Qty	10.000

From the “Qty_list” column we get the total quantity and the total number of Unique Items and compare with the the extracted data from this part of the PDF:

Total Items / Qty : 5 / 10.000

Conclusion

In conclusion, the Invoice Data Extraction System successfully meets the outlined requirements, delivering a reliable and accurate(99.3056%) solution for extracting and validating invoice data from various PDF formats. The system's performance(0.06 seconds/PDF), scalability, and robustness contribute to its effectiveness in real-world applications.