PDF to XML Conversión

Eshita Shukla - VJTI
Jui Pitale - VJTI
Aditya Tumarada - IIT Guwahati

Introduction

Problem Statement

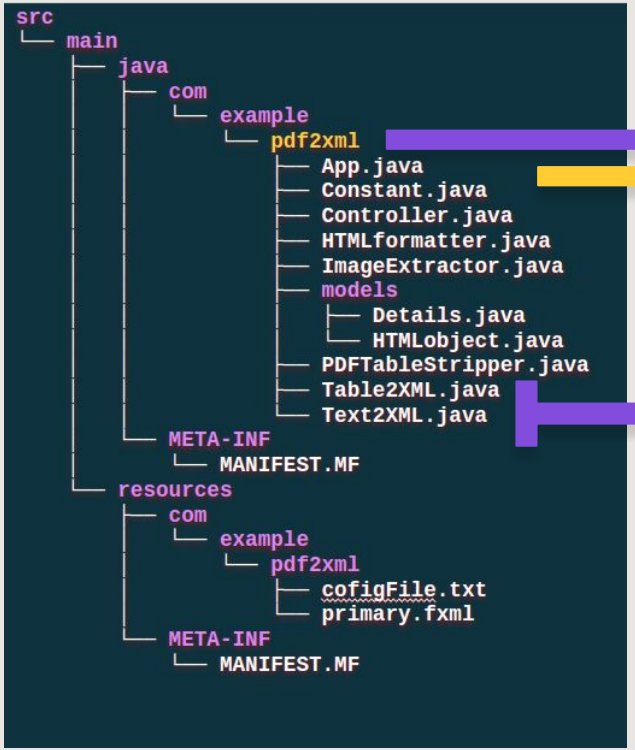
Write a PDF to XML utility (tool) by leveraging the pdfbox library so that we can use this tool to compare .pdf files to DB tables.

We have a framework, which can compare the XML data with database tables and reports the data differences.

However, the system does not support the data comparison between PDF and table data due to which data is compared manually and leads to error-prone.

So that once we convert the PDF file to the XML file, it can be injected into the existing Framework to compare the data and find the root cause the data discrepancy in the PDF files.

Project Structure



Location of **all classes**

main Class

Classes for **generation of .xml**

Demo

Approach

Our Target

Implementation with GUI

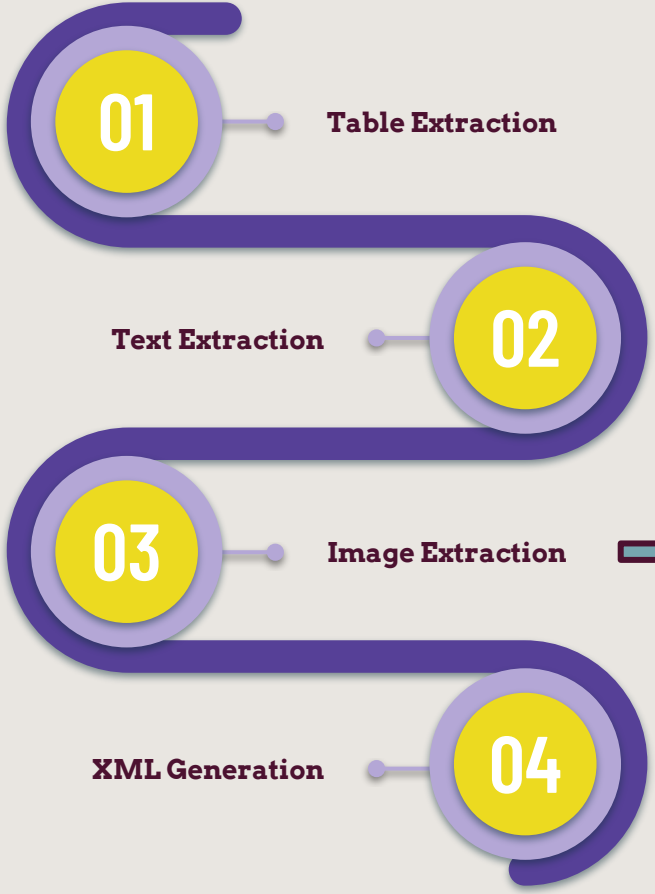
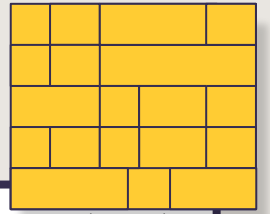


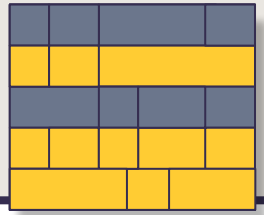
Table Extraction

Extract rows & their coordinates

Divide each page into rectangles

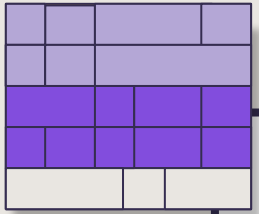


Divide each rectangle (row) into columns



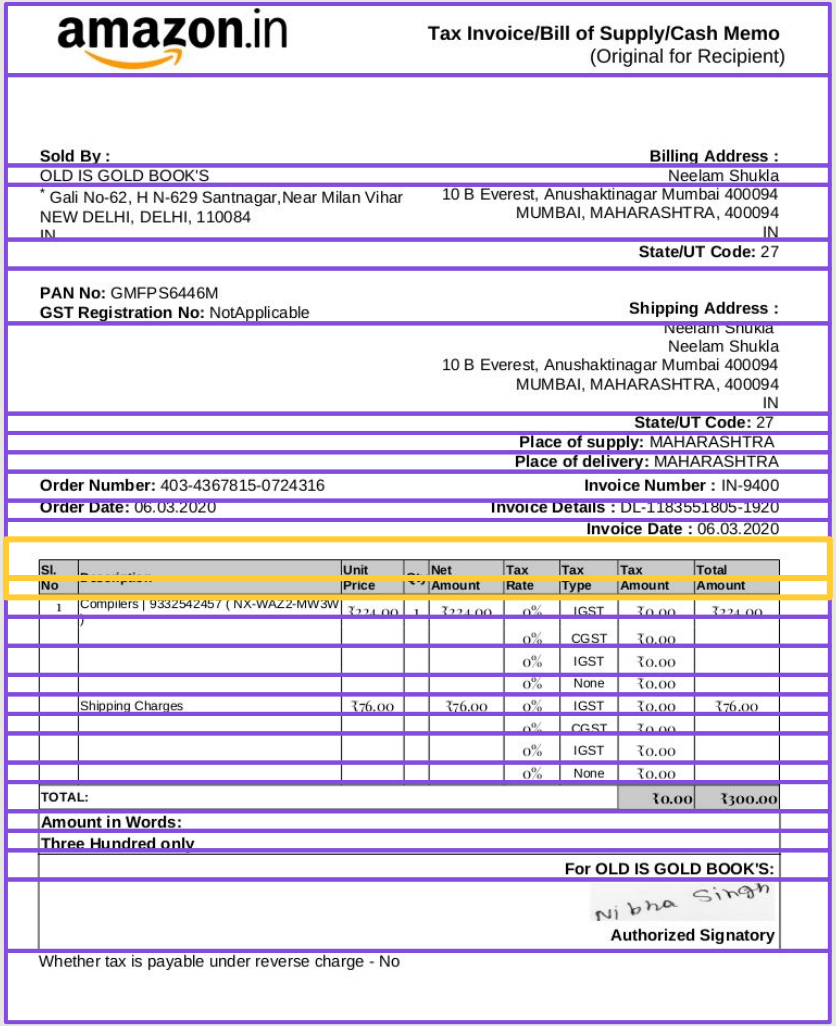
Classify each row ad "normal" or "heading"

Group all rows into table (some in no tables)



Group all rows into table (some in no tables)

Dividing into rectangles



amazon.in

Tax Invoice/Bill of Supply/Cash Memo
(Original for Recipient)

Sold By :
OLD IS GOLD BOOK'S
* Gali No-62, H N-629 Santnagar,Near Milan Vihar
NEW DELHI, DELHI, 110084
IN

Billing Address :
Neelam Shukla
10 B Everest, Anushaktinagar Mumbai 400094
MUMBAI, MAHARASHTRA, 400094
IN
State/UT Code: 27

PAN No: GMFPS6446M
GST Registration No: NotApplicable

Shipping Address :
Neelam Shukla
Neelam Shukla
10 B Everest, Anushaktinagar Mumbai 400094
MUMBAI, MAHARASHTRA, 400094
IN
State/UT Code: 27

Place of supply: MAHARASHTRA
Place of delivery: MAHARASHTRA

Order Number: 403-4367815-0724316
Order Date: 06.03.2020

Invoice Number : IN-9400
Invoice Details : DL-1183551805-1920
Invoice Date : 06.03.2020

Sl. No	Description	Unit Price	Qty	Net Amount	Tax Rate	Tax Type	Tax Amount	Total Amount
1	Compilers | 9332542457 (NX-WAZ2-MW3W)	₹224.00	1	₹224.00	0%	IGST	₹0.00	₹224.00
					0%	CGST	₹0.00	
					0%	IGST	₹0.00	
					0%	None	₹0.00	
	Shipping Charges	₹76.00		₹76.00	0%	IGST	₹0.00	₹76.00
					0%	CGST	₹0.00	
					0%	IGST	₹0.00	
					0%	None	₹0.00	
TOTAL:							₹0.00	₹300.00

Amount in Words:
Three Hundred only

For OLD IS GOLD BOOK'S:

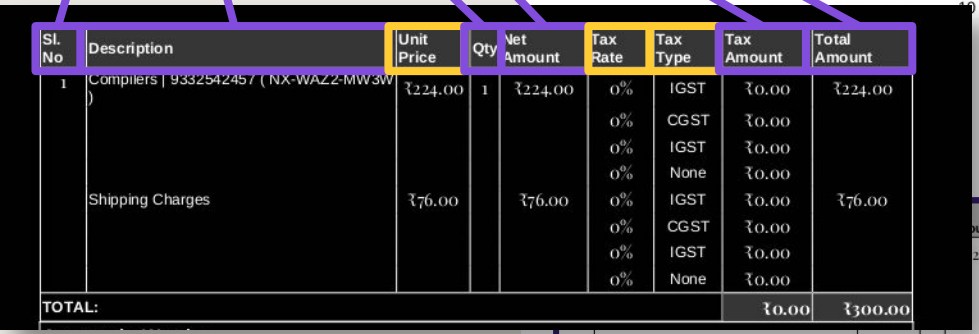
Nibha Singh

Authorized Signatory

Whether tax is payable under reverse charge - No

Classifying into "heading" or not

Probability = 6/9



amazon.in

Tax Invoice/Bill of Supply/Cash Memo
(Original for Recipient)

Sold By :
OLD IS GOLD BOOK'S
* Gali No-62, H N-629 Santnagar,Near Milan Vihar
NEW DELHI, DELHI, 110084
IN

Billing Address :
Neelam Shukla
10 B Everest, Anushaktinagar Mumbai 400094
MUMBAI, MAHARASHTRA, 400094
IN
State/UT Code: 27

PAN No: GMFPS0446M
GST Registration No.: NotApplicable

Shipping Address :
Neelam Shukla
Neelam Shukla
10 B Everest, Anushaktinagar Mumbai 400094
MUMBAI, MAHARASHTRA, 400094
IN
State/UT Code: 27
Place of supply: MAHARASHTRA
Place of delivery: MAHARASHTRA
Invoice Number : IN-9400
Invoice Details : DL-1183551805-1920
Invoice Date : 06.03.2020

Sl. No	Description	Unit Price	Qty	Net Amount	Tax Rate	Tax Type	Tax Amount	Total Amount
1	Compilers | 9332542457 (NX-WAZ2-MW3W)	₹224.00	1	₹224.00	0%	IGST	₹0.00	₹224.00
					0%	CGST	₹0.00	
					0%	IGST	₹0.00	
					0%	None	₹0.00	
	Shipping Charges	₹76.00		₹76.00	0%	IGST	₹0.00	₹76.00
					0%	CGST	₹0.00	
					0%	IGST	₹0.00	
					0%	None	₹0.00	
TOTAL:							₹0.00	₹300.00

Amount in Words:
Three Hundred only

For OLD IS GOLD BOOK'S:
Nibha Singh
Authorized Signatory

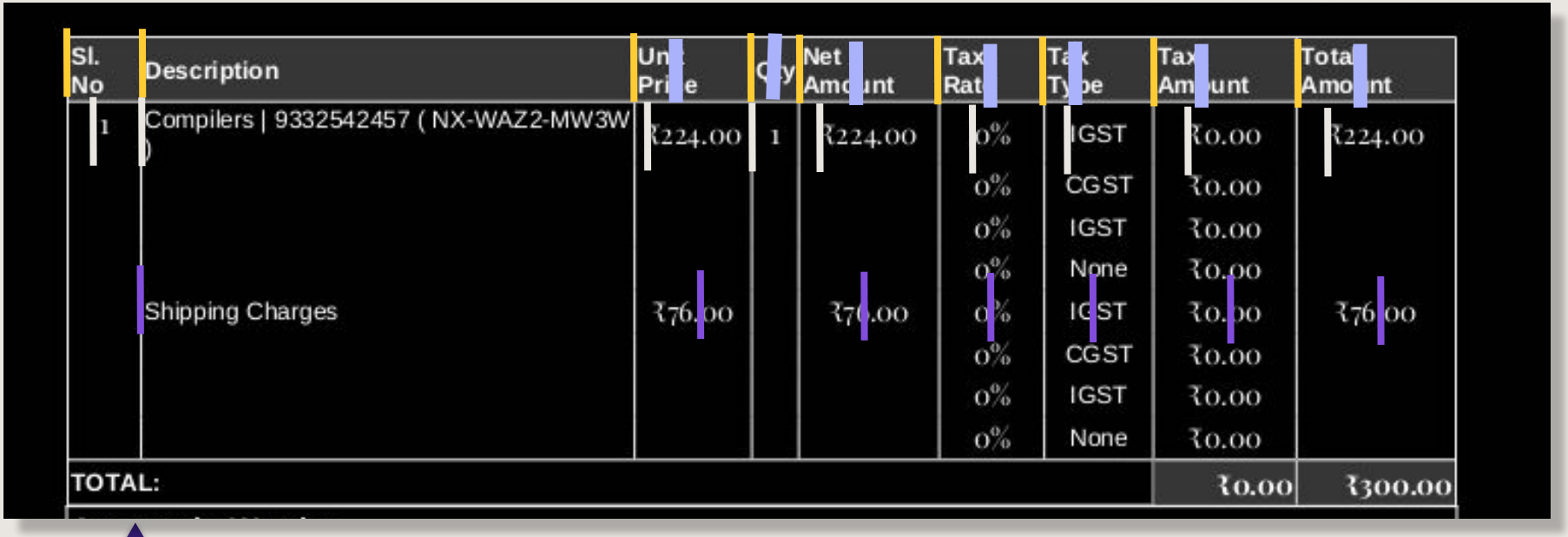
Whether tax is payable under reverse charge - No

configFile

```
Description
DESCRIPTION
 No.
Qty
Amount
AMOUNT
```

Without probability calculation:
if the keyword "Description" is present in the name of the book that has been bought by the customer, the algo fails

Probability Calculation for Consecutive rows



Sl. No	Description	Unit Price	Qty	Net Amount	Tax Rate	Tax Type	Tax Amount	Total Amount
1	Compilers | 9332542457 (NX-WAZ2-MW3W)	₹224.00	1	₹224.00	0%	IGST	₹0.00	₹224.00
					0%	CGST	₹0.00	
					0%	IGST	₹0.00	
					0%	None	₹0.00	
	Shipping Charges	₹76.00		₹76.00	0%	IGST	₹0.00	₹76.00
					0%	CGST	₹0.00	
					0%	IGST	₹0.00	
					0%	None	₹0.00	
TOTAL:							₹0.00	₹300.00

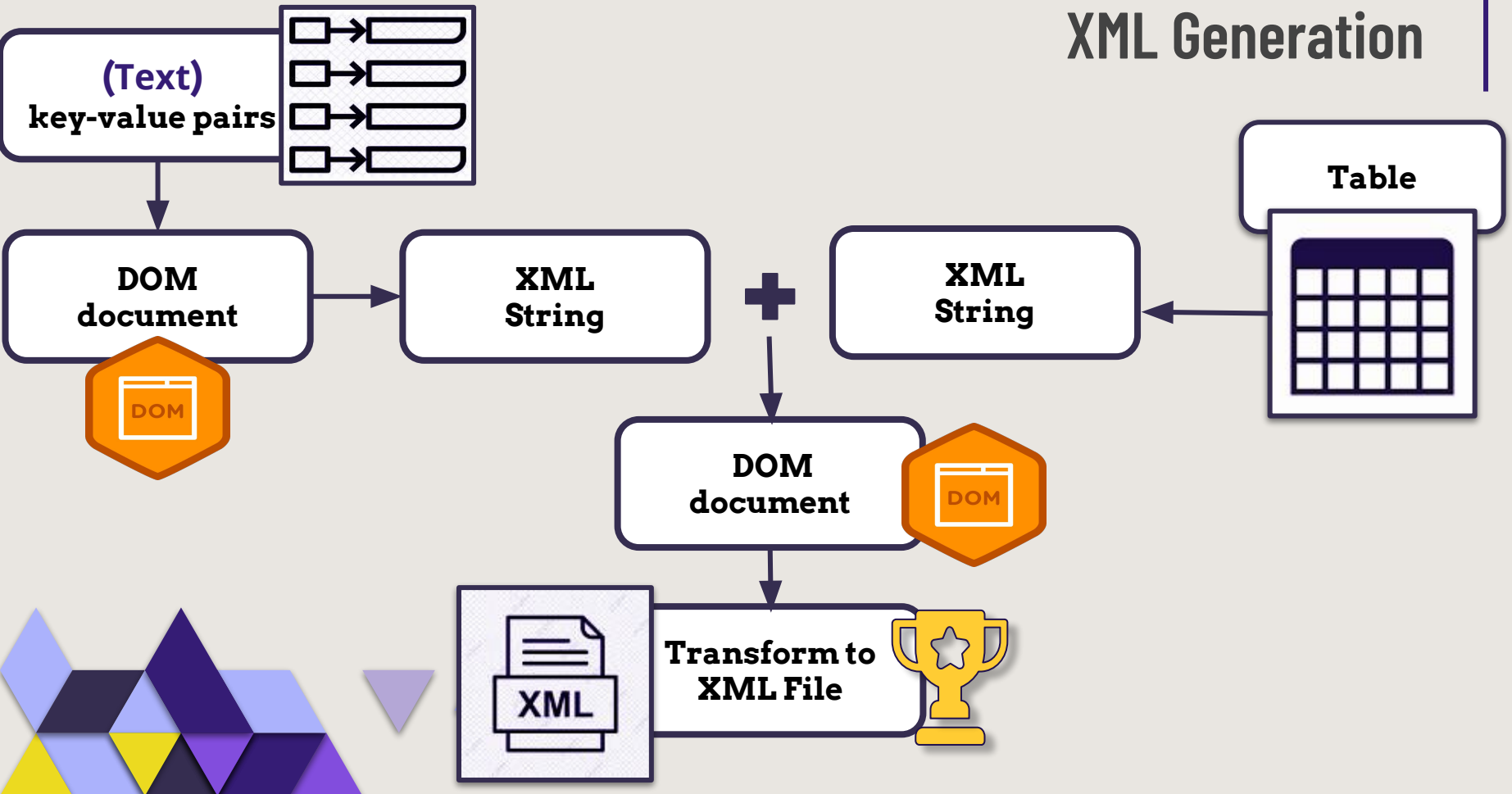
APPROACH

Text Extraction

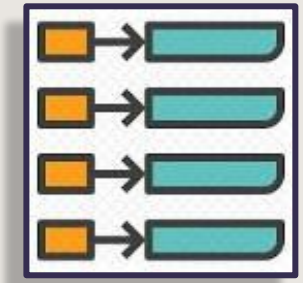
(page-wise extraction in key value pair form)

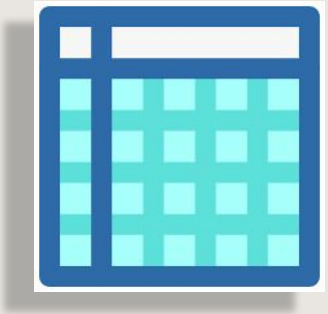
APPROACH

XML Generation



Limitations of the S/W





Recognition of the "key"-"value" pairs

- Change in **text styles**
- **Line-spacing**
- **Could be difficult** to handle

At least some of the **words** in the **heading** should be known

Conclusión

- Our team has been working on this problem statement for the last **5 weeks**.
- We have been able to process **almost all types of PDFs** and successfully extracted all data into an xml, to ensure which we changed our algorithm thrice.
- On the **basis of only the alignment**(left/right/center) of the text in each column of the table, we are able to **successfully get the desired XML output.**
- The project is made as **flexible** as possible. The user can add more keywords if they want

THANKS!

Questions ??


THANKS FOR YOUR ATTENTION
DO YOU HAVE ANY QUESTIONS?
memecrunch.com