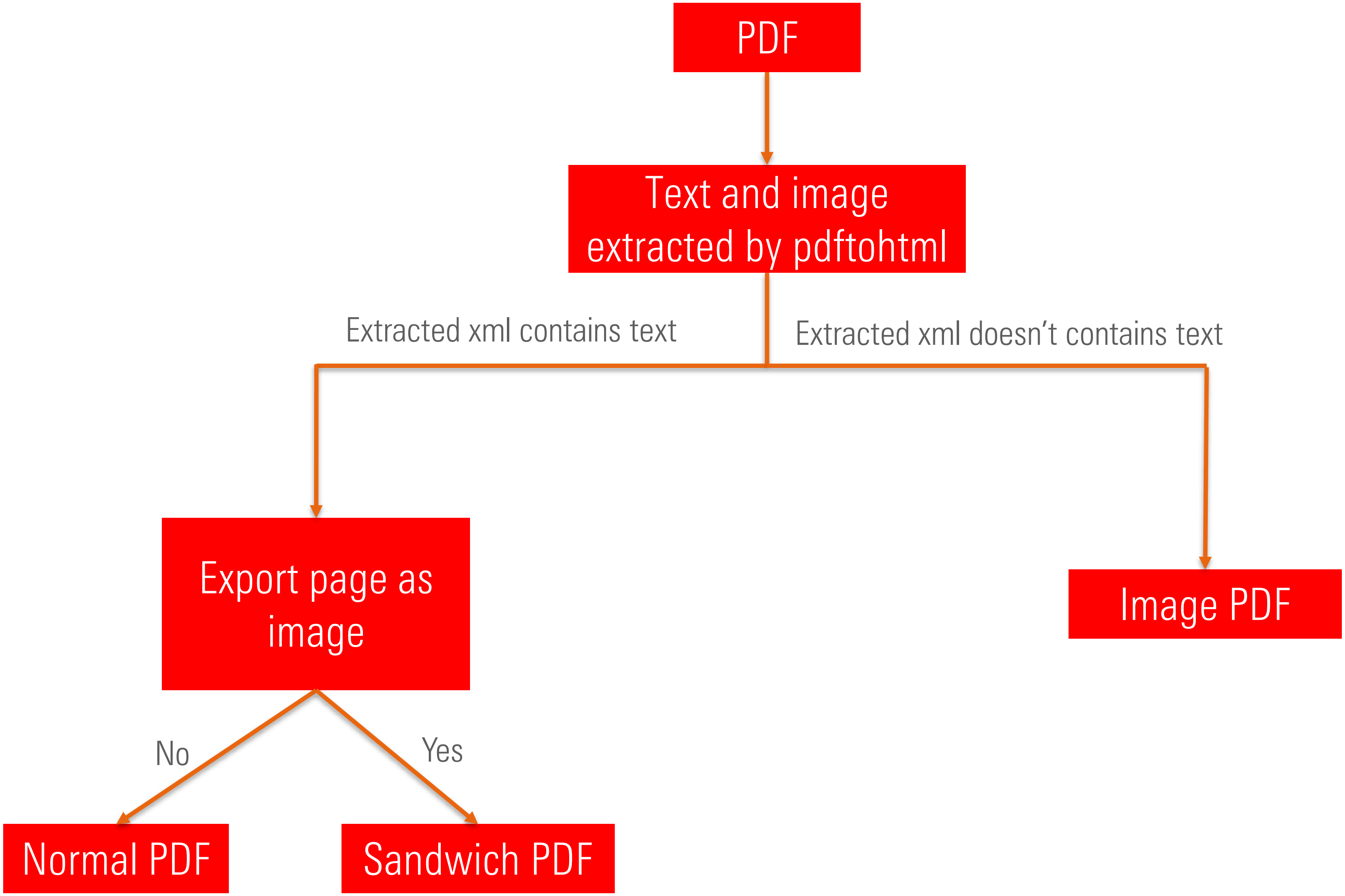
PDF Table Extractor

-Pratik Bhavsar

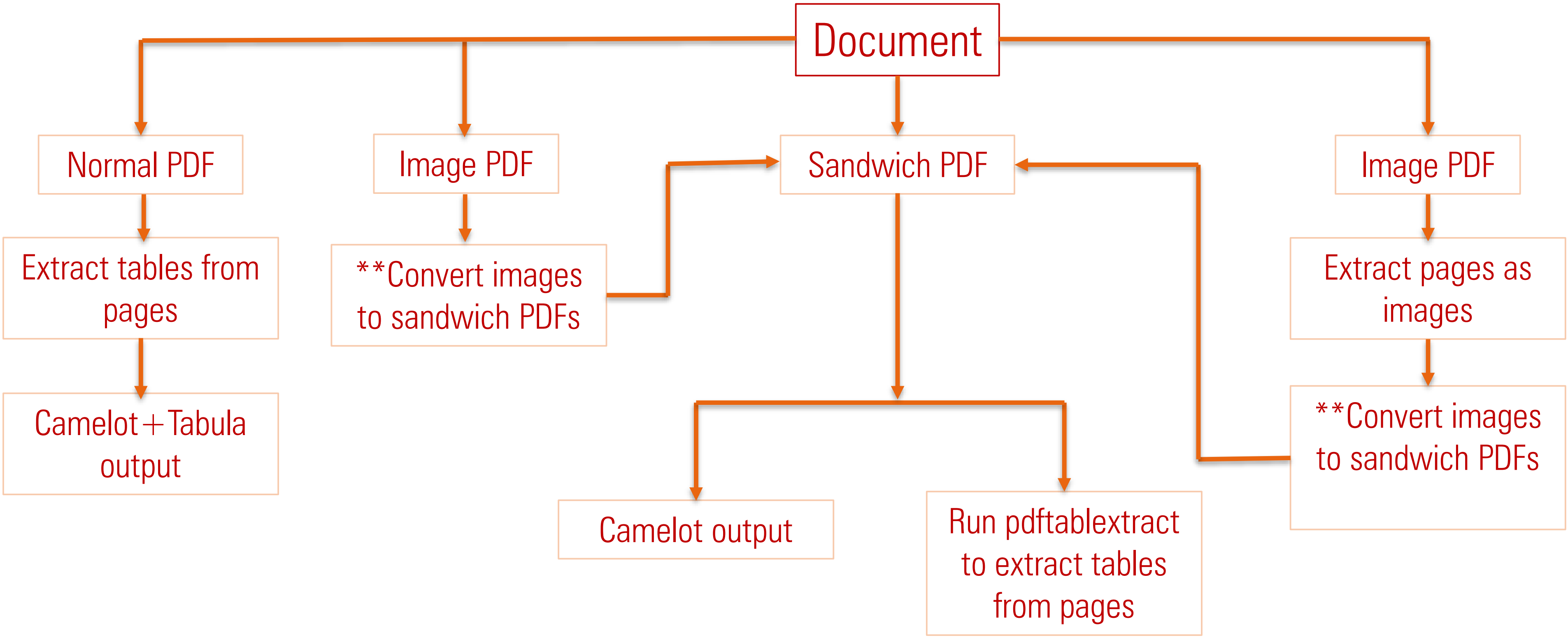
How it works!

Find type of PDF



PDF
Text and image extracted by pdftohtml
Extracted xml contains text
Extracted xml doesn’t contains text
Export page as image
Image PDF
No
Yes
Normal PDF
Sandwich PDF

Flow chart



Document

Normal PDF

Extract tables from pages

Camelot+Tabula output

Image PDF

**Convert images to sandwich PDFs

Sandwich PDF

Camelot output

Run pdftablextract to extract tables from pages

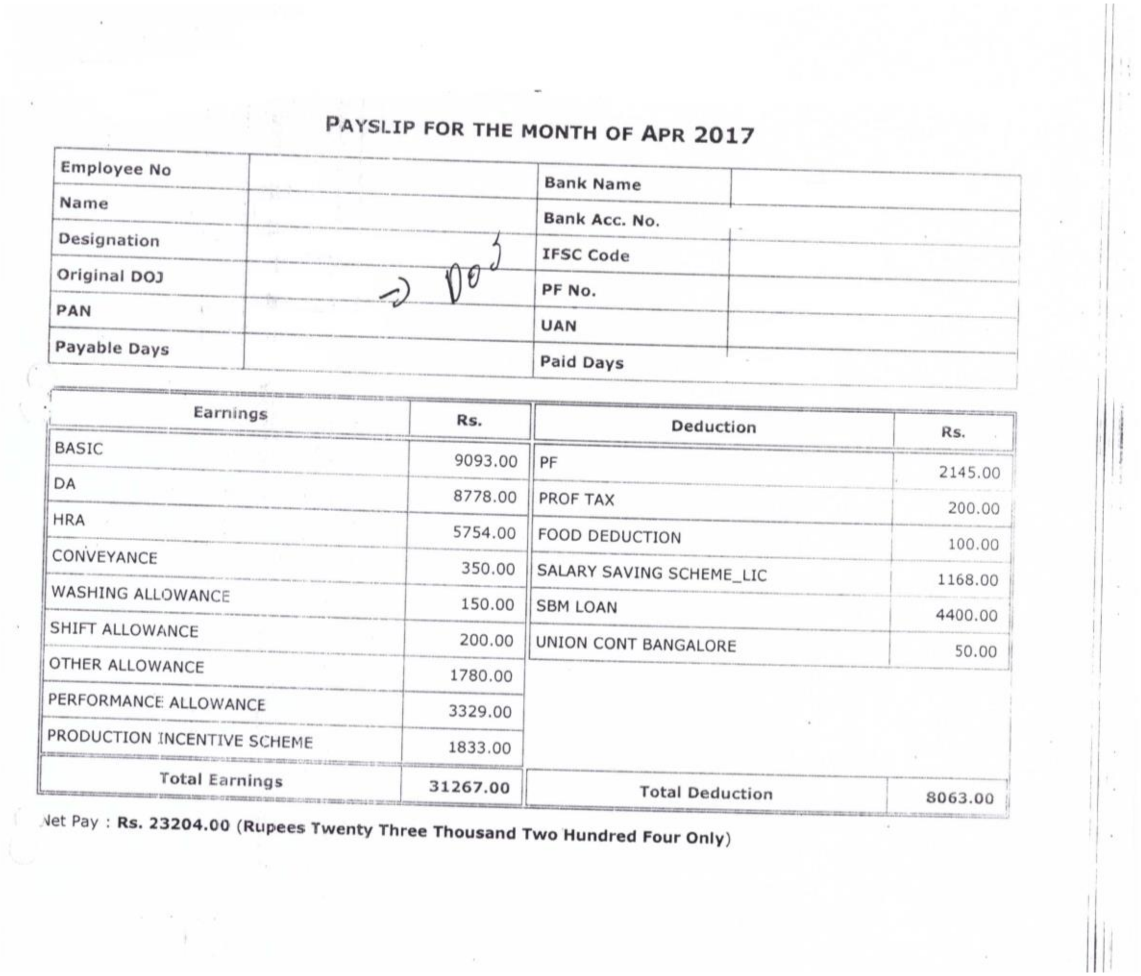
Image PDF

Extract pages as images

**Convert images to sandwich PDFs

**Processing intensive and hence done on multi-threading

- pdftablextract



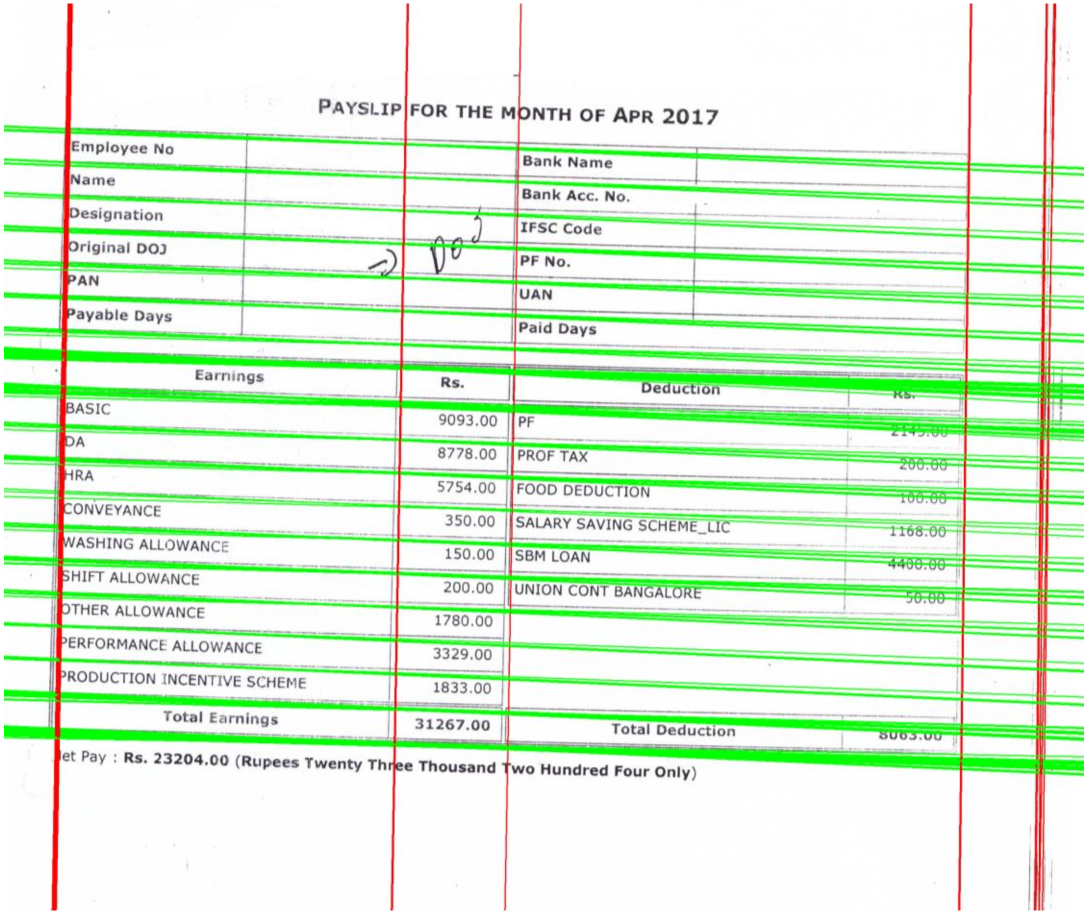
PAYSLIP FOR THE MONTH OF APR 2017

Employee No		Bank Name	
Name		Bank Acc. No.	
Designation		IFSC Code	
Original DOJ	→ DOJ	PF No.	
PAN		UAN	
Payable Days		Paid Days	

Earnings	Rs.	Deduction	Rs.
BASIC	9093.00	PF	2145.00
DA	8778.00	PROF TAX	200.00
HRA	5754.00	FOOD DEDUCTION	100.00
CONVEYANCE	350.00	SALARY SAVING SCHEME_LIC	1168.00
WASHING ALLOWANCE	150.00	SBM LOAN	4400.00
SHIFT ALLOWANCE	200.00	UNION CONT BANGALORE	50.00
OTHER ALLOWANCE	1780.00		
PERFORMANCE ALLOWANCE	3329.00		
PRODUCTION INCENTIVE SCHEME	1833.00		
Total Earnings	**31267.00**	**Total Deduction**	**8063.00**

Net Pay : **Rs. 23204.00 (Rupees Twenty Three Thousand Two Hundred Four Only)**

- pdftablextract



PAYSLIP FOR THE MONTH OF APR 2017

Employee No		Bank Name	
Name		Bank Acc. No.	
Designation		IFSC Code	
Original DOJ	→ Doj	PF No.	
PAN		UAN	
Payable Days		Paid Days	

Earnings	Rs.	Deduction	Rs.
BASIC	9093.00	PF	2145.00
DA	8778.00	PROF TAX	200.00
HRA	5754.00	FOOD DEDUCTION	100.00
CONVEYANCE	350.00	SALARY SAVING SCHEME_LIC	1168.00
WASHING ALLOWANCE	150.00	SBM LOAN	4400.00
SHIFT ALLOWANCE	200.00	UNION CONT BANGALORE	50.00
OTHER ALLOWANCE	1780.00		
PERFORMANCE ALLOWANCE	3329.00		
PRODUCTION INCENTIVE SCHEME	1833.00		
Total Earnings	31267.00	Total Deduction	8063.00

et Pay : **Rs. 23204.00** (**Rupees Twenty Three Thousand Two Hundred Four Only**)

Output folder structure

Put the image/PDF to be parsed in test folder and the outputs can be found in outputs folder in a folder with the name of image/PDF

- /output
  - original_filename-camelot.csv
  - original_filename-tabula.csv
  - original_filename-tablextract.csv

Output folder structure

Individual file output folders

- /tables – contains our required tables in 2 folders
  - /camelot – Outputs by camelot library which is preferred
  - /tabula – outputs by tabula library
  - Tables extracted by pdftablextract
- /texts – contains temporary text
- /images – contains intermediate image exports
- /logs – contains logs for debugging

Config

```
{

    "input_folder": "../test",

    "output_folder": "../outputs",

    "delete_temp": “true",

    “pool_size": 8,

    "generate_images_dpi": 200,

    "MIN_COL_WIDTH": 60,

    "MIN_ROW_WIDTH": 60

}
```

Solution highlights

- Works on all kinds of PDFs and images
- Uses all open source libraries
- Uses multithreading for faster processing
- Solution made in Python

Dataset referred

Given dataset and a collection of pdfs of annual report of companies

Current challenges

- Problems in images/image based pdf
  - Difficult to detect table grid in tables without lines
  - Difficult to find table grid in pages containing both table and text
  - Difficulty with images with small font or low DPI

Thank you.