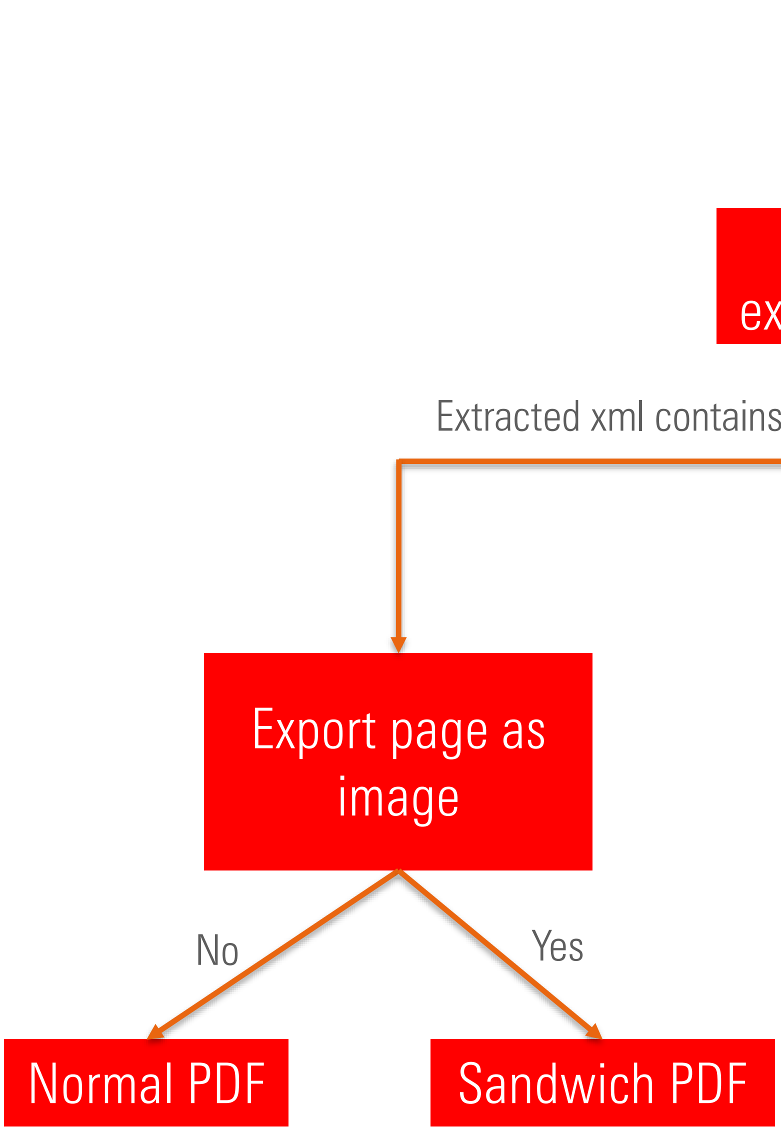
PDF Table Extractor

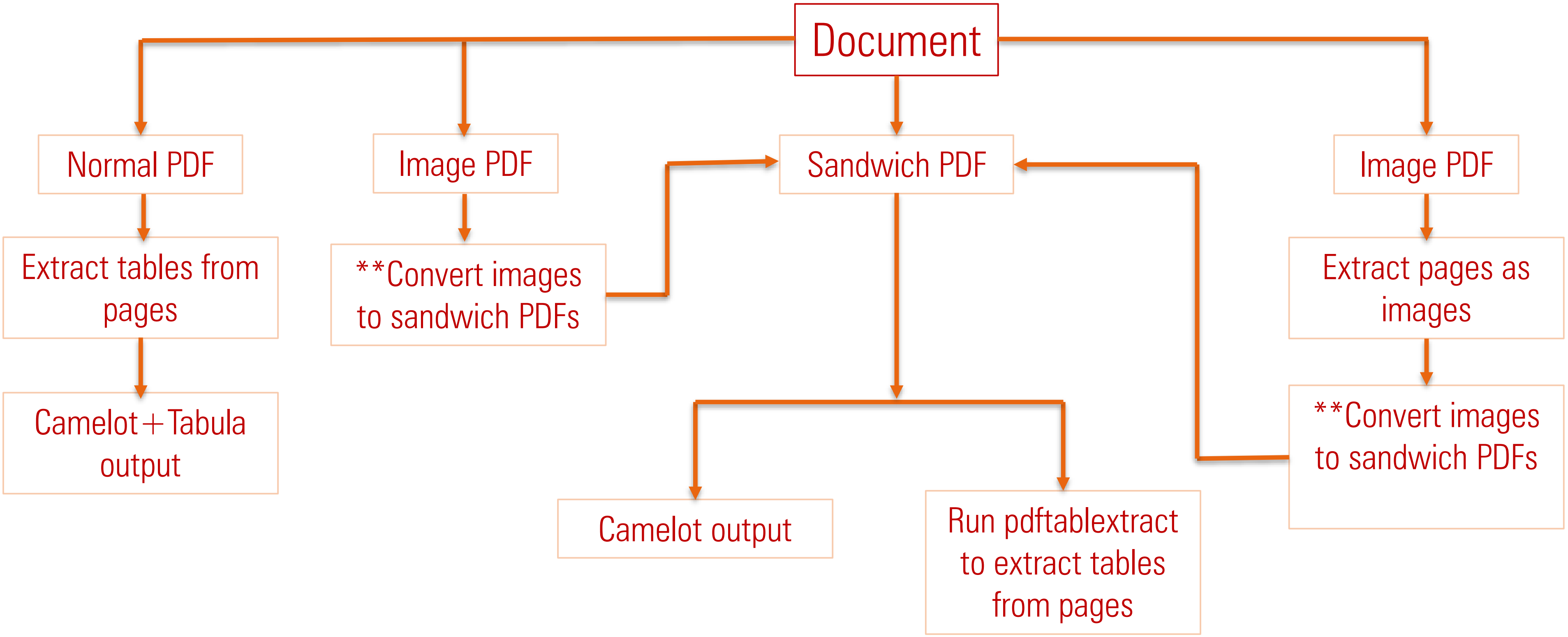
-Pratik Bhavsar

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
Image PDF
Sandwich PDF
Image PDF
Extract tables from pages
**Convert images to sandwich PDFs
Extract pages as images
Camelot+Tabula output
**Convert images to sandwich PDFs
Camelot output
Run pdftablextract to extract tables from pages

**Processing intensive and hence done on multi-threading

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 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
- 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
  - Images with small font or low DPI

Demo – Please check the demo video.

Thank you.