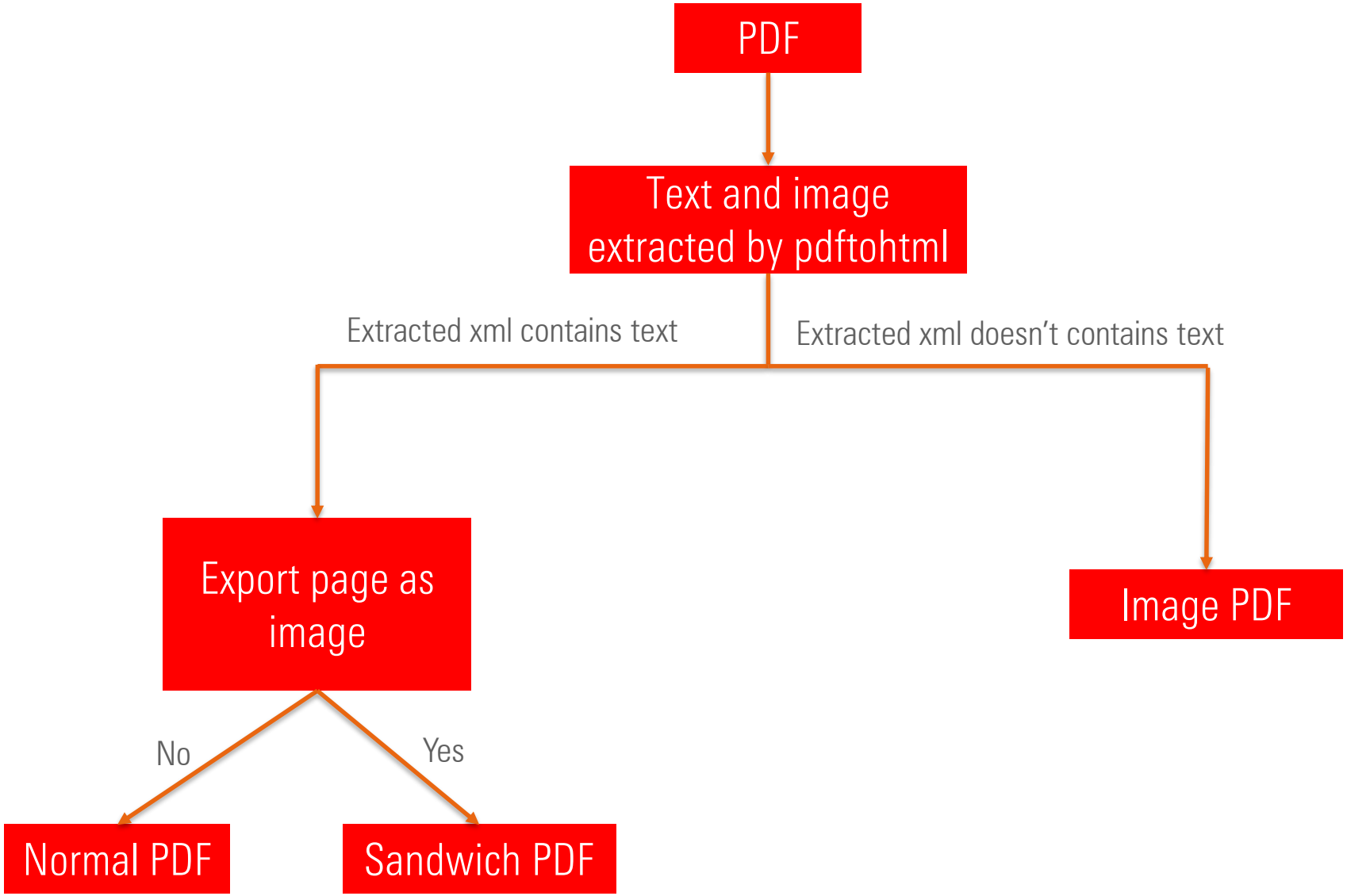
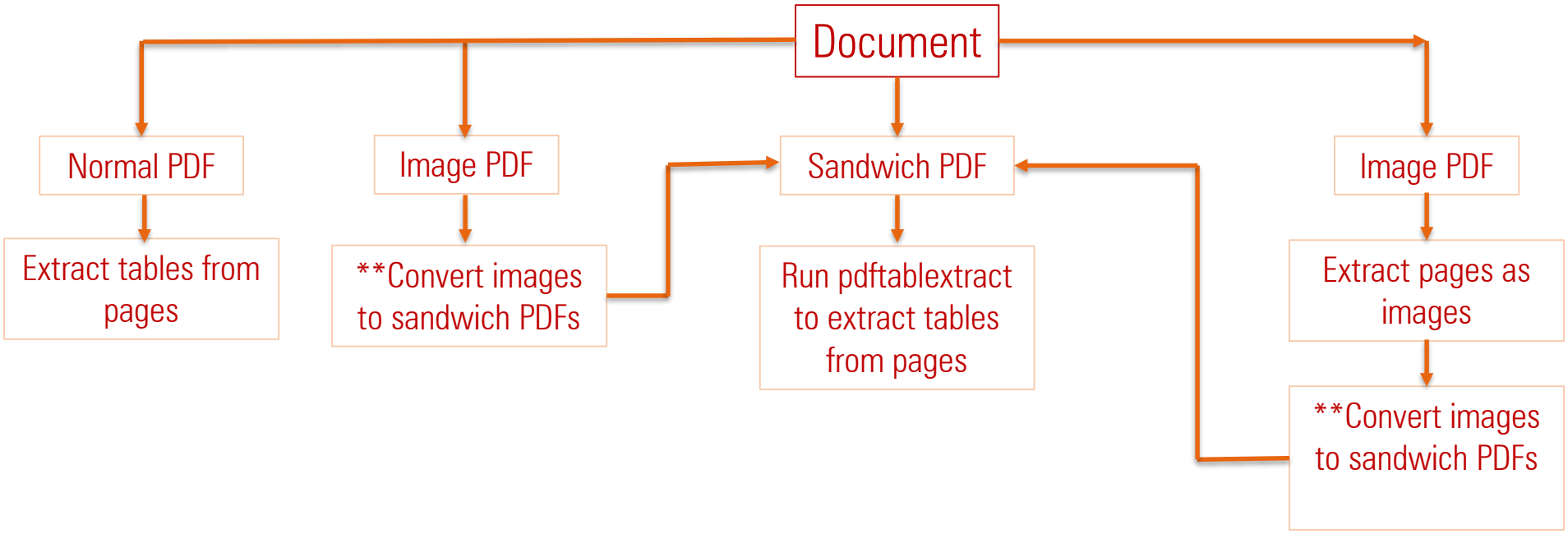
PDF Table Extractor

-Pratik Bhavsar

Find type of PDF



Flow chart



**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

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

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
- Requires no manual intervention to work
- Uses all open source libraries
- Solution made in Python which has become de-facto for NLP and image processing problems

Demo – Please check the demo video.

Thank you.