

1. OCR Scanned Document

As per provided sample document. The layout is standardized as header, footer, body, Logo name area,

Information can be identifiable in different regions.

Cropping the image to a certain region of Interest improves the document parsing in the following ways.

1. Cropping to specific region externally provide a line break for generalized string matching. Otherwise, a specific regular expression is required for matching the content. Then it won't follow a common pattern.
2. Logos may have multiple lines of content. Logos usually don't follow patterns. So, selecting a logo region can extract its content without interference from another line character.
3. In left and right alignment of text could be a problem when detecting line by line.
4. The line of content will be disturbed by external user entries (stamps, written text) because they don't follow horizontal lines.
5. And these kinds of documents don't have vertical and horizontal line separators in tables. So, the layout identification of tables is not great. So available table-py models won't work with this type of pdf.

2. Packages and Libraries

```
pip install pytesseract pdf2image opencv-python poppler
```

These are the following libraries I used.

3. Tesseract installer for Windows

installation needed to complete using the following link otherwise it will throw an error:

<https://github.com/UB-Mannheim/tesseract/wiki>

4. To Run this:

1. Install Tesseract
2. Set Tesseract path
3. Set Poppler path
4. Set pdf_folder path
 - a. You can check the structure inside the submission folder

5. Layout Selection

Region of Interest 1: Top Right

Region of Interest for name

Region of Interest 2: Top Right

Tan Tock Seng HOSPITAL
National Healthcare Group
The TSH Community Fund is a charitable trust set up to support the needy patients and the community through financial assistance. To donate or know more about TSH Community Fund, please visit www.tsh.com.sg/tshfund. Thank You.

No. 11 Jalan Tan Tock Seng, Singapore 308443
 Tel: 6256 6011 (Main Line), 6357 7000 (Nurse Appointment Room), 6511 4338 (Billing Enquiries)
 Fax: 6256 9234 Reg No. 199003683N

TAX INVOICE
(Adjusted)

TO: MDM. BLK # SINGAPORE

PATIENT NAME : dsdasdaasd

00829
Follow up claim

MRN/NRIC : S- A
 CASE NO : 12183607201-00021
 VISIT DATE : 11.06.2018 08:30
 LOCATION : TCT5A
 INVOICE DATE : 08.07.2018
 TYPE OF SUPPLY : CASH/CREDIT
 GST REG NO : M2-0094564-6

PLEASE PAY UPON RECEIPT OF THIS INVOICE

Description	Amount(S\$)
Total Amount Payable	353.23
ADJUSTMENT:	
ROUND DOWN FOR AMOUNT PAYABLE BY PATIENT	0.02
PAYMENT:	
INTEGRATED GREAT EASTERN SUPREMEHEALTH	0.00
MEDISAVE	318.04
TOTAL DUE AFTER PAYMENT	15.30
DUE FROM:	
INTEGRATED GREAT EASTERN SUPREMEHEALTH	15.30
MEDISAVE	0.00
	0.00
FOR INFORMATION	
INTEGRATED GREAT EASTERN SUPREMEHEALTH payout consist of the following:	
MEDISHIELD LIFE	304.25
GREAT EASTERN SUPREMEHEALTH ADDITIONAL COVERAGE	13.79
For more information on the payment details, please contact Customer Service Hotline Customer Service at 1800 248 2686	

MEDISAVE A/C HOLDER

CPF NO

Amt Deducted

S. A 20.02

Total amount payable after GST is \$378.14.
 Total GST for this bill at 7% is \$24.76 which is absorbed by the Government.
 The amount payable by patient has been rounded down to the nearest 5 cents.

Region of Interest Name and page

Region of Interest Footer

6. Assumptions

- I assumed the discharged date is the bill received to date from Rubber seal. Due to its format of DD-MMM-YYYY

7. JSON Formatting

- In making JSON data, I followed to add data into an array of key values. Not included as key-value pair information.
- This Document contains lots of special characters. And also, the Tesseract finds some low dpi content as special content. But as per matching regex, most of the extracted information is useful.
 - Eg -> Document contains “:” semicolon but OCR detects it as “-” but this “-” can’t be replicable because some fields use “-” to represent values.
 - In data extraction generalizing this text cleaning will disturb other values
 - Names and contents are different in both document layouts.