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

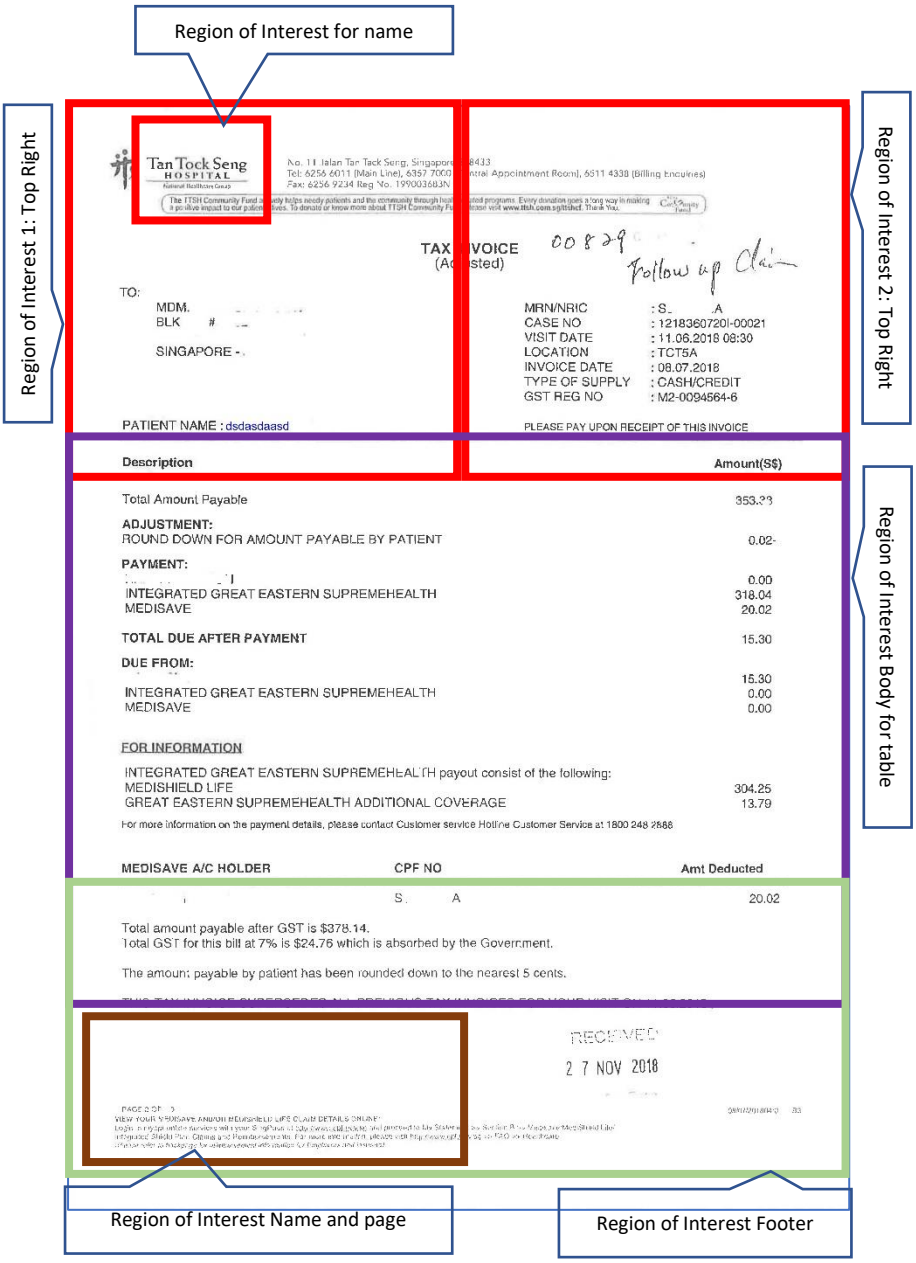
installation needed to complete using the following link otherwise it will throw an error:

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

4. To Run this:

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

5. Layout Selection



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.