Task

Information from the provided medical bill needs to be extracted and formatted into JSON (Check **Sample JSON out** below) format. These data can be header information (Key-value pair), table formatted data or in free form (Page number etc). You can check the expected values from the **Expected Fields** section below.

You need to select proper methodology for key value extraction and table data extraction. For this you are free to use existing open source models/ solutions and combine them into the application. The proposed solution needs to be generalized so that it will extract information from similar types of document.

Submissions for this assignment (Script) can be in the format of a Notebook (Jupyter/Colab) or Python Files via GitHub. Clearly mention the libraries and any installation information in ReadMe. If any part is not clear , you can take your own assumptions while developing the solution and clearly mention the assumptions while submitting the response.

Medical bill : Link

Expected Fields

Following information need to be extracted from the documents (Only if available) :

1) Header Information

- Patient Name
- Patient National ID
- Hospital / clinic
- GST Reg No of the hospital/clinic [if available])
- Visit Date
- Tax Invoice Date
- Bill /receipt date
- Admission date
- Discharge date of the invoice
- Tax invoice number/Invoice No of the invoice/receipt),
- Bill reference number
- Total Amount Payable
- Invoice Page (invoice page as shown on the restructured hospital bill, e.g., 1 OF 2 and 2 OF 2)
- Doctor Name
- Location
- Bill Type [ORIGINAL/DUPLICATE/INTERIM]

2) Line-Item Information (Table)

- Description
- Quantity
- Amount (Line item's amount),
- Code
- Date (Date of the line item, this field is available in some invoices)

Sample JSON out

Header items and Line items (Table) from each page need to be extracted separately like follows :


```
[
  {
    "Page_Number":1,
    "Table":[
      {
        "DESCRIPTION":"FOUNDATION ONE SENT TO USA",
        "UNIT PRICE":"8,250.00",
        "QTY":"1.00",
        "TOTAL":"8,250.00"
      },
       {
        "DESCRIPTION":"Consultation",
        "UNIT PRICE":"1,250.00",
        "QTY":"2.00",
        "TOTAL":"2500.00"
      }

    ],
    "Key_Values":{
      "BillRefNumber":"04D60",
      "TaxInvoiceDate":"7.000.00",
      "Patient":"19,60",
      "TypeofSupply":"28,200.85",
      "PatientNRICHRN":"17.25",
      "PaymentClass":"24.20",
      "VisitDate":"18,200.85"
    }
  },
  {
    "Page_Number":2,
    "Table":[
      {
        "DESCRIPTION":"FOUNDATION ONE SENT TO USA",
        "UNIT PRICE":"8,250.00",
```

```
        "QTY":"1.00",
        "TOTAL":"8,250.00"
      },
      {
        "DESCRIPTION":"Consultation",
        "UNIT PRICE":"1,250.00",
        "QTY":"2.00",
        "TOTAL":"2500.00"
      }
    ],
    "Key_Values":{
      "BillRefNumber":"04D60",
      "TaxInvoiceDate":"7.000.00",
      "Patient":"19,60",
      "TypeofSupply":"28,200.85",
      "PatientNRICHRN":"17.25",
      "PaymentClass":"24.20",
      "VisitDate":"18,200.85"
    }
  }
]
```


Note -

1) You are open to use any available solution/git repos/libraries for this task.
   Eg : Paddle OCR , Layout detection, Table detection

2) Keep in mind that there are no ‘right answers.’ This assignment is designed to gauge your skills and give us an idea of how you approach tasks relevant to the role. Complete the assignment as much as you can.