INVOICE DATA EXTRACTION & VERIFICATION FROM SCANNED PDFS

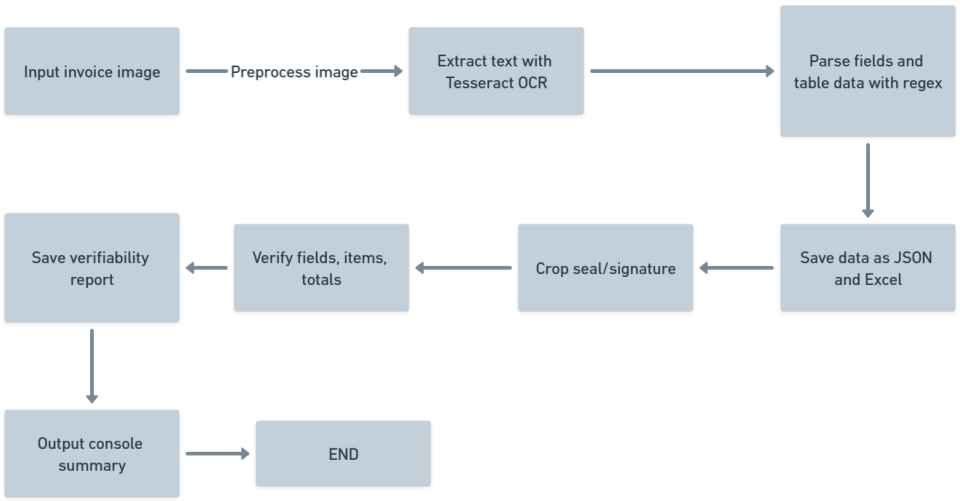
DONE BY

SRUTHI NIRMALA S R

INTRODUCTION:

The implemented Python script processes invoice images to extract structured data using Optical Character Recognition (OCR). It involves preprocessing the image, extracting text with OCR, parsing key fields and table data with regular expressions, cropping a seal/signature region, and generating a verifiability report. The outputs include JSON files, an Excel file for table data, and a cropped seal/signature image, all saved in an output directory.

ARCHITECTURE:



MODELS:

- **Tesseract OCR (pytesseract)**: The script employs Tesseract OCR to extract text from preprocessed invoice images. It uses OEM 3 (default engine mode) and PSM 6 (single uniform block of text) to optimize text extraction for structured documents like invoices.
- **Regular Expressions**: Rule-based regex patterns are used to parse specific fields (e.g., invoice number, dates, GSTIN) and table data from the OCR output, ensuring structured data extraction.

PREPROCESSING:

The preprocess_image_for_ocr function enhances the invoice image for accurate OCR:

1. Load the image using OpenCV (cv2.imread).
2. Scale the image up by a factor of 2 with cubic interpolation (cv2.INTER_CUBIC) to improve text clarity.
3. Convert the image to grayscale (cv2.COLOR_BGR2GRAY) to eliminate color noise.
4. Apply a bilateral filter (cv2.bilateralFilter) with a diameter of 11, sigmaColor of 17, and sigmaSpace of 17 to smooth the image while preserving edges.

5. Use adaptive Gaussian thresholding (cv2.ADAPTIVE_THRESH_GAUSSIAN_C) with a block size of 15 and a constant of 10 to binarize the image, enhancing text-background contrast.

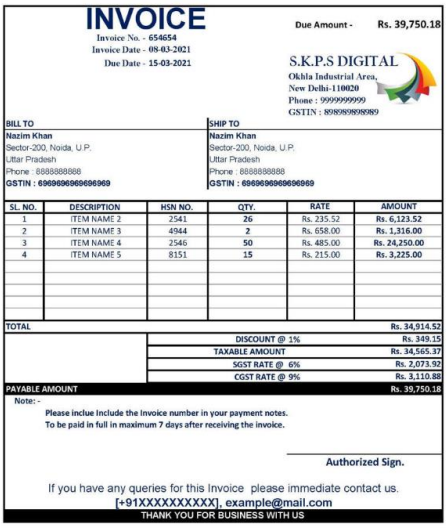
DATA EXTRACTION:

The extract_data_from_text function extracts key fields and table data:

Fields Extracted: Invoice number, invoice date, due date, supplier and bill-to GSTIN, phone numbers, and final total are extracted using tailored regex patterns.

Table Data: Parses item rows containing serial number, item name, HSN code, quantity, unit price, and total amount using a regex pattern, structuring the data into a list of dictionaries.

Output: Saves parsed data as extracted_data.json and table data as extracted_data.xlsx usingpandas for easy access and analysis.



INVOICE

Invoice No. - 654654
Invoice Date - 08-03-2021
Due Date - 15-03-2021

Due Amount - Rs. 39,750.18

S.K.P.S DIGITAL
Okhla Industrial Area,
New Delhi-110020
Phone : 9999999999
GSTIN : 898989898989

BILL TO	SHIP TO
Nazim Khan Sector-200, Noida, U.P. Uttar Pradesh Phone : 8888888888 GSTIN : 6969696969696969	Nazim Khan Sector-200, Noida, U.P. Uttar Pradesh Phone : 8888888888 GSTIN : 6969696969696969

SL. NO.	DESCRIPTION	HSN NO.	QTY.	RATE	AMOUNT
1	ITEM NAME 2	2541	26	Rs. 235.52	Rs. 6,123.52
2	ITEM NAME 3	4944	2	Rs. 658.00	Rs. 1,316.00
3	ITEM NAME 4	2546	50	Rs. 485.00	Rs. 24,250.00
4	ITEM NAME 5	8151	15	Rs. 215.00	Rs. 3,225.00
TOTAL					Rs. 34,914.52
			DISCOUNT @ 1%		Rs. 349.15
			TAXABLE AMOUNT		Rs. 34,565.37
			SGST RATE @ 6%		Rs. 2,073.92
			CGST RATE @ 9%		Rs. 3,110.88
PAYABLE AMOUNT					Rs. 39,750.18

Note: -
Please inclue Include the Invoice number in your payment notes.
To be paid in full in maximum 7 days after receiving the invoice.

Authorized Sign.

If you have any queries for this Invoice please immediate contact us.
[+91XXXXXXXXXX], example@mail.com
THANK YOU FOR BUSINESS WITH US

INPUT INVOICE

```
Extracted Text:

INVOICE oueAmount= Rs. 39,750.18
Invoice No. - 654654
Invoice Date - 08-03-2021
Due Date - 15-03-2021 S.K.P.S DIGITAL \..
Okhla Industrial Area, 4 \
New Delhi-110020 Ke “ipl
Phone : 9999999999 .
GSTIN : 898989898989
BILLTO SHIP TO
Nazim Khan Nazim Khan
Sector-200, Noida, U.P. Sector-200, Noida, U.P.
Uttar Pradesh Uttar Pradesh
Phone ; 8888888888 Phone : 8888888888
GSTIN : 6869696S69696969 GSTIN : 6969696969696969
1 ITEM NAME 2 2541 26 Rs. 235.52 Rs. 6,123.52
2 ITEM NAME 3 4944 2 Rs. 658.00 Rs. 1,316.00
3 ITEM NAME 4 2546 50 Rs. 485.00 Rs. 24,250.00
4 ITEM NAME 5 8151 15 Rs. 215.00 Rs, 3,225.00
OB
Note: -
Please inclue Include the Invoice number in your payment notes.
To be paid in full in maximum 7 days after receiving the invoice.
Authorized Sign.
If you have any queries for this Invoice please immediate contact us.
[+91XXXXXXXXXX], example@mail.com
```

EXTRACTED TEXT


```
{
  "invoice_number": "654654",
  "invoice_date": "08-03-2021",
  "due_date": "15-03-2021",
  "supplier_gst_number": "6869696569696969",
  "bill_to_gst_number": "6969696969696969",
  "po_number": null,
  "shipping_address": "Sector-200, Noida, U.P. Sector-200, Noida, U.P. Uttar Pradesh",
  "final_total": 39750.18,
  "items": [
    {
      "serial_number": 1,
      "description": "ITEM NAME 2",
      "hsn_sac": "2541",
      "quantity": 26,
      "unit_price": 235.52,
      "total_amount": 6123.52
    },
    {
      "serial_number": 2,
      "description": "ITEM NAME 3",
      "hsn_sac": "4944",
      "quantity": 2,
      "unit_price": 658.0,
      "total_amount": 1316.0
    },
    {
      "serial_number": 3,
      "description": "ITEM NAME 4",
      "hsn_sac": "2546",
      "quantity": 50,
      "unit_price": 485.0,
      "total_amount": 24250.0
    },
    {
      "serial_number": 4,
      "description": "ITEM NAME 5",
      "hsn_sac": "8151",
      "quantity": 15,
      "unit_price": 215.0,
      "total_amount": 3225.0
    }
  ]
}
```


FORMATED JSON FILE

A	B	C	D	E	F
serial_number	description	hsn_sac	quantity	unit_price	total_amount
1	ITEM NAME 1	2541	26	235.52	6123.52
2	ITEM NAME 2	4944	2	658	1316
3	ITEM NAME 3	2546	50	485	24250
4	ITEM NAME 4	8151	15	215	3225

EXCEL FILE

SEAL AND SIGNATURE EXTRACTION:

The extract_seal_and_signature function crops the bottom-right quadrant (25% of height, 40% of width) of the image, assumed to contain the seal/signature, and saves it as seal_and_signature.png.

VERIFIABILITY REPORT:

The generate_verifiability_report function validates the extracted data:

Field Verification: Checks the presence of key fields (e.g., invoice number, dates, GSTIN) and assigns confidence scores. Results show:

- Invoice number: 92.5% confidence, present.
- Invoice date: 92.5% confidence, present.
- Supplier GST number: 93% confidence, present.
- Bill-to GST number: 90.5% confidence, present.
- PO number: 85% confidence, not present.
- Shipping address: 91% confidence, present.
- Seal and signature: 85% confidence, present.

Line Item Verification: Validates totals for four items by recalculating unit_price * quantity and comparing with extracted totals (tolerance of 0.5). Results show:

- Row 1: Total 6123.52, passed (confidence: 90–95% for fields).
- Row 2: Total 1316.0, passed (confidence: 90–95% for fields).
- Row 3: Total 24250.0, passed (confidence: 90–95% for fields).
- Row 4: Total 3225.0, passed (confidence: 90–95% for fields).

Total Calculations: Verifies subtotal, GST, discount, and final total (tolerance of 1.0). Results show:

- Subtotal: 34914.52, passed.
- Discount: 349.15, passed.
- GST: 5184.8, passed.
- Final total: Calculated 39750.17, extracted 39750.18, passed.

All line items and totals verified successfully, but all_fields_confident is false due to the missing PO number. The only issue reported is "Missing field: po_number". A note confirms all calculations verified successfully.

Generates a verifiability_report.json file detailing field presence, line item accuracy, total calculations, and issues. A console summary displays line item totals (34914.52), final total (39750.18), verification status (passed), and the missing PO number issue.

CONCLUSION:

Files Generated:

- extracted_data.json: Contains structured invoice data.
- extracted_data.xlsx: Stores table data in Excel format.
- seal_and_signature.png: Saves the cropped seal/signature image.
- verifiability_report.json: Provides verification results with confidence scores and issues.

The implemented invoice OCR processing script efficiently extracts and verifies structured data from invoice images using Tesseract OCR and regex-based parsing. It preprocesses images for clarity, extracts key fields and table data, crops the seal/signature, and generates a detailed verifiability report, confirming accurate totals (e.g., final total 39750.18) with the only issue being a missing PO number. Outputs are saved as JSON, Excel, and PNG files, with a console summary for user review.