Subiksha K

CB.SC.I5DAS21161

YAVAR- Internship assignment

**Title:** Invoice Data Extraction & Verification from Scanned PDFs

**Approach 2:** Enhanced Invoice Data Extraction with Confidence Scoring & Streamlit UI

1. PDF to Image Conversion - PyMuPDF (fitz)
2. Optical Character Recognition (OCR)
3. Information Extraction - Regex patterns
4. Confidence Estimation
5. Final Total Calculation - Auto-calculated Final Total = Subtotal - Discount
6. Streamlit Deployment

Requirements & Installation

Python Libraries:

* pytesseract – OCR engine wrapper
* PyMuPDF – PDF to image conversion
* pandas – Data handling + Excel export
* Pillow – Image handling
* streamlit – For UI
* re – Built-in module for regex matching

External Installations:

* Tesseract-OCR executable (installed locally and path set)
* VS Code + Virtual Environment for isolated and clean development setup

Output:

* 1. Text extraction in a raw .txt file:

Superstore INVOICE

# 37425

Date: Oct 24 2012

Bill To: Ship To: .

Ship Mode: First Class

Aaron Hawkins 94109, San

Francisco, California, Balance Due: $160.94

United States

Item Quantity Rate Amount

GBC VeloBind Cover Sets 4 $49.41 $197.63

Binders, Office Supplies, OFF-BI-4372

Subtotal: $197.63

Discount (20%) : $39.53

Shipping: $2.84

Total: $160.94

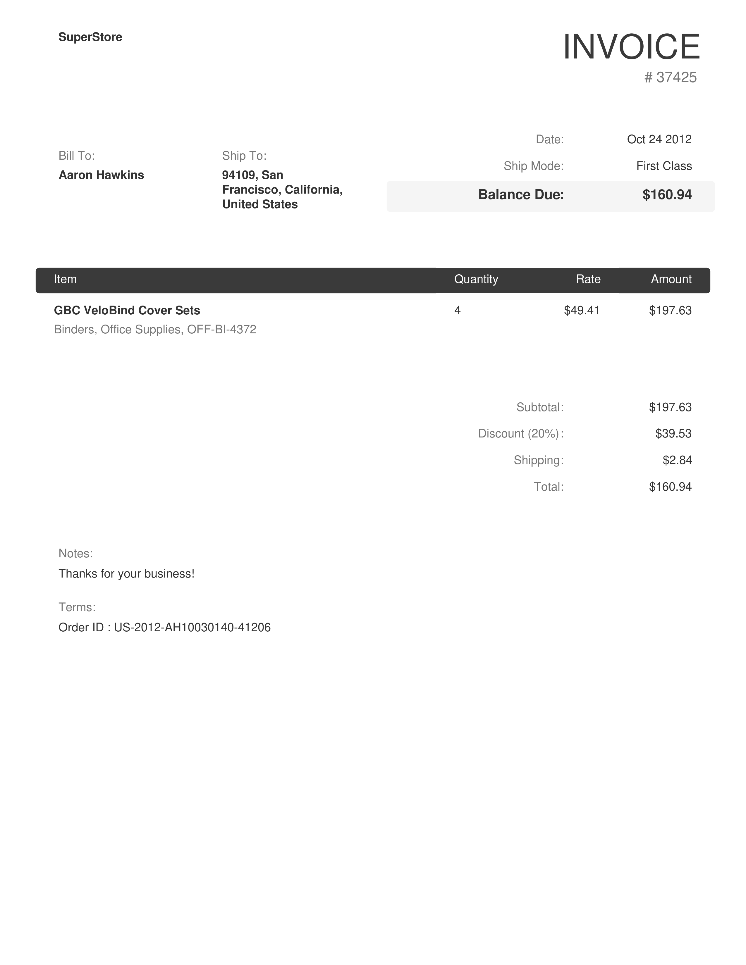
Notes:

Thanks for your business!

Terms:

Order ID : US-2012-AH10030140-41206

* 1. pdf to image converted



* 1. Excel output:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Invoice Number** | **Date** | **Customer** | **Ship Mode** | **Balance Due** | **Subtotal** | **Discount** | **Shipping** | **Total** | **Order ID** | **Final Total** |
| 37425 | Oct 24 2012 |  | First Class | 160.94 | 197.63 | 39.53 | 2.84 | 160.94 | US-2012-AH10030140-41206 | 158.1 |

* 1. Json output:  
     {

  "Invoice Number": {

    "value": "37425",

    "confidence": 95

  },

  "Date": {

    "value": "Oct 24 2012",

    "confidence": 96

  },

  "Customer": {

    "value": "Ship To: .",

    "confidence": 96

  },

  "Ship Mode": {

    "value": "First Class",

    "confidence": 96

  },

  "Balance Due": {

    "value": "160.94",

    "confidence": 0

  },

  "Subtotal": {

    "value": "197.63",

    "confidence": 0

  },

  "Discount": {

    "value": "39.53",

    "confidence": 0

  },

  "Shipping": {

    "value": "2.84",

    "confidence": 0

  },

  "Total": {

    "value": "160.94",

    "confidence": 0

  },

  "Order ID": {

    "value": "US-2012-AH10030140-41206",

    "confidence": 69

  },

  "Final Total": {

    "value": "158.1",

    "confidence": 100

  }

}

**Demo:**



Inference:   
It takes files from the input folder and gives outputs in JSON and Excel formats, along with a basic confidence report.

The JSON and Excel outputs were generated correctly, but the confidence scores mostly show zero — even for fields that were actually extracted well. This happened because I haven’t implemented a proper way to calculate or match confidence with reference data yet. Also, the seal and signature cropping part is still pending.

Though the output isn't fully complete, this version lays the groundwork for offline invoice extraction and gives a good starting point for improvements.