Subiksha K

CB.SC.I5DAS21161

YAVAR- Internship assignment

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

**Approach 1:**

1. PDF to Image Conversion
2. Optical Character Recognition (OCR)
3. Text Normalization (preprocessing)
4. Rule-Based Information Extraction
5. Data Storage

**Models Used**

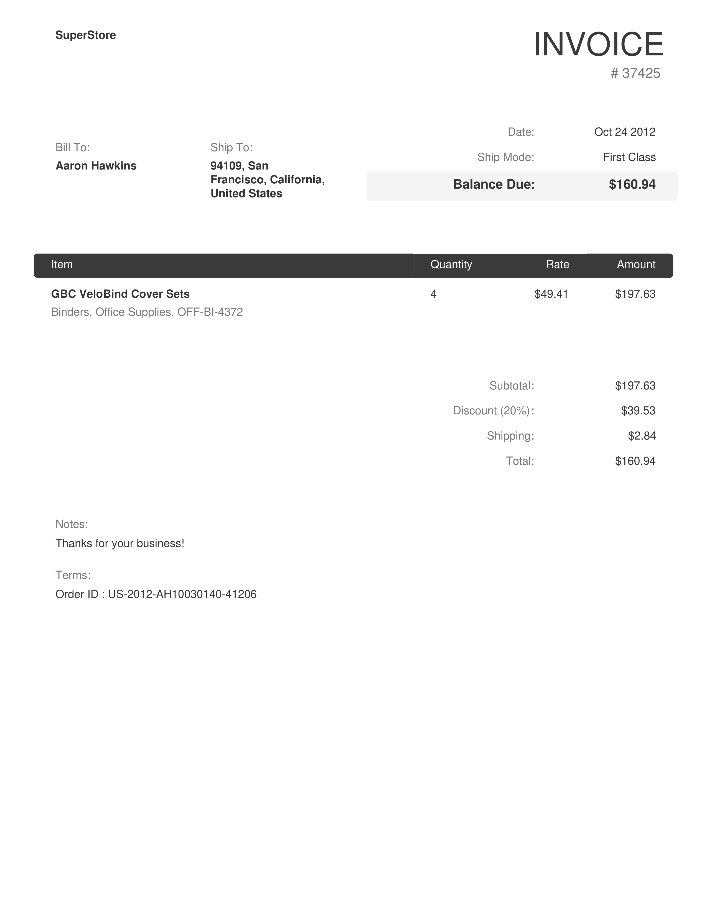
* **Tesseract OCR Engine**: This open-source OCR engine (accessed via the pytesseract Python wrapper) is responsible for converting the pixel-based images into textual data.
* **Regular Expressions (re module)**: Rule-based "model" for pattern matching and extracting specific data points from the OCR'd text.
* **Requirements and Installation**

To run this project, the following components were installed:

* **Python Libraries**:
  + pytesseract: Python wrapper for Tesseract.
  + pandas: For data manipulation and Excel export.
  + pdf2image: For converting PDFs to images.
  + Pillow (PIL): For image processing.
  + Tesseract executable and poppler was installed

Output:

Pdf to png converted output:



.json output:  
{

    "invoice\_id": "37425",

    "date": "Oct\n24\n2012",

    "bill\_to": "",

    "ship\_to": "",

    "item\_name": "",

    "rate": "49.41",

    "amount": "197.63",

    "subtotal": "197.63",

    "discount": "39.53",

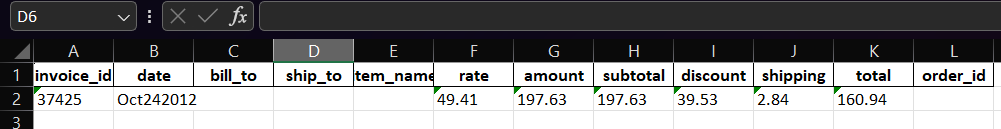
    "shipping": "2.84",

    "total": "160.94",

    "order\_id": ""

}

.xlxs output:



INFERENCE:

OCR significantly underperformed, failing to extract key fields (Bill To, Ship To, Item Name, Order ID) despite accurate numerical capture.