OCR Tool for Parsing Invoices and Receipts

Ilayda Gülyaz (22108086), Rohan Mitra (12300406), Katerina Aleksova (21109556), Moumita Basu (12304782)

Computer Vison

6.7.2024

Agenda

1. Project Goal
2. Dataset
3. Framework
4. Conclusion
5. Demo

Project Goal

Building an Optical Character Recognition (OCR) Tool for Parsing Invoices and Receipts

Motivation

- hard-copy documents digitalized
  - users can search, format, edit
  - shareable
- quickly lookup numbers, addresses, names

Advantages

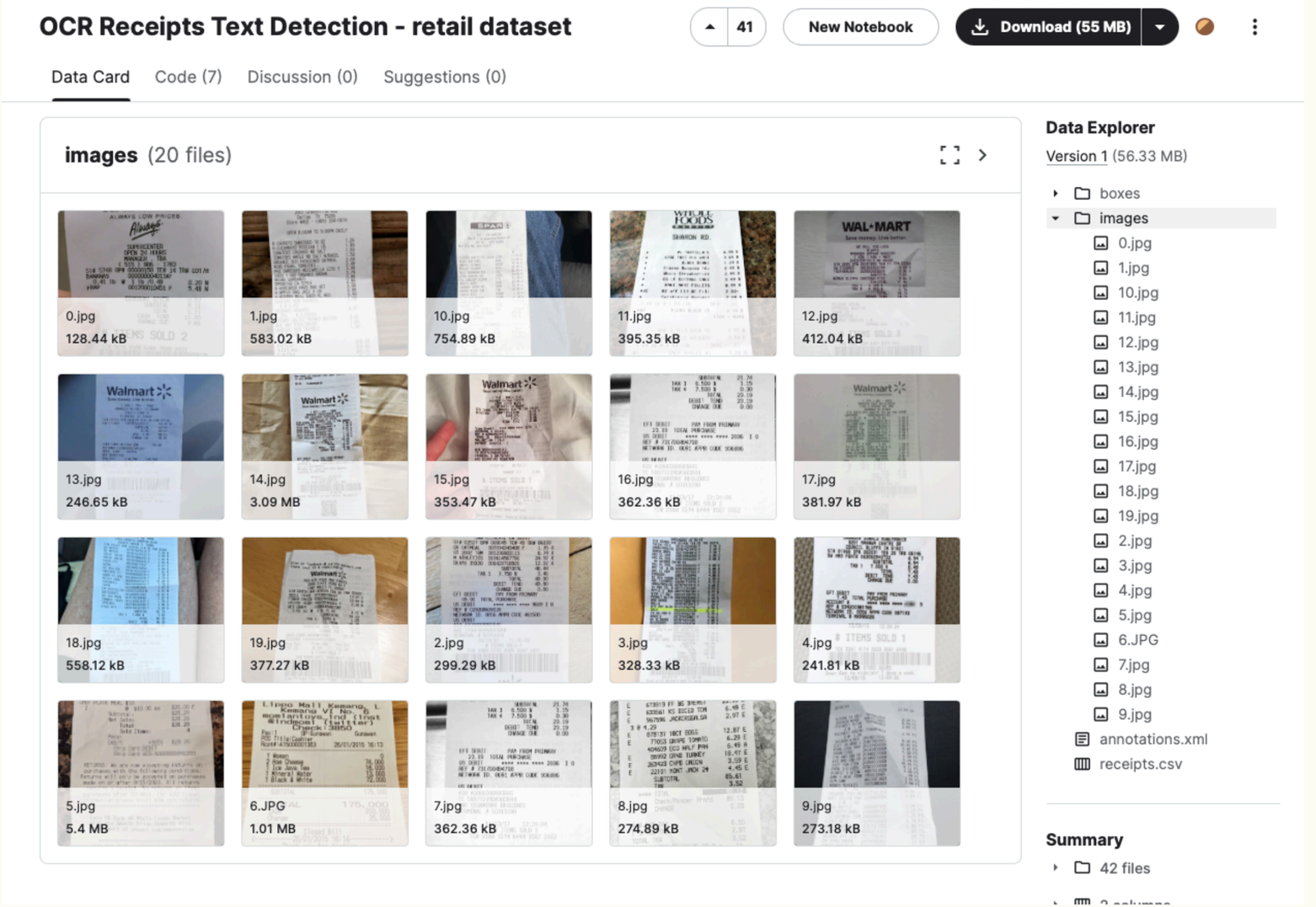
- Minimize & eliminate errors in data entry
- no need for manual data entry & physical storage
- Save labor costs --> automated
- different file formats

Dataset

OCR Receipts Text Detection - retail dataset found on Kaggle

https://www.kaggle.com/datasets/trainingdatapro/ocr-receipts-text-detection/data

It is a collection of photos captured from various **grocery store receipts**. This dataset is specifically designed for tasks related to **Optical Character Recognition (OCR)**.



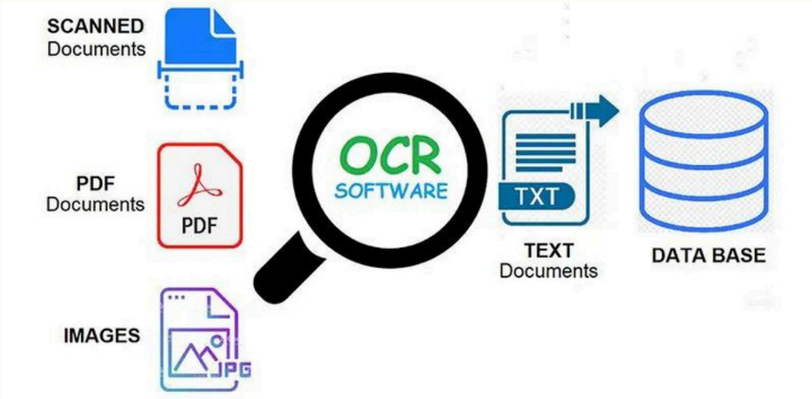
Framework: Backend Overview

Flask

The backend of the project is built using Flask, a lightweight and flexible web framework.

Key Features:

1. **Image Upload**
2. **OCR Processing:**
   - Utilizes **Tesseract** to perform OCR on uploaded files.
   - Uses **OpenCV** for image preprocessing to enhance OCR accuracy.
3. **Table Extraction** (structured DataFrames for easy data manipulation)
4. **Text Parsing** (item details, prices, quantities...)



OpenCV

python

Tesseract OCR

2024-06-22T19-07_export

	Item	Quantity	Unit	Price	SKU	Store Name	Purchase Date
0	rcarruts shredded	10	oz	1.29		(TRADER JOE'S	06-28-2014
1	rcucumbers persian lb	1		4.99		(TRADER JOE'S	06-28-2014
2	tomatoes crushed no salt	1		1.59		(TRADER JOE'S	06-28-2014
3	tomatoes whole no salt wbasil	1		1.59		(TRADER JOE'S	06-28-2014
4	organic oldfashiuned oatheal	1		2.69		(TRADER JOE'S	06-28-2014
5	pkg shredded mozzarella lite t	1		3.99		(TRADER JOE'S	06-28-2014
6	eggs organic brown	1	doz	3.79		(TRADER JOE'S	06-28-2014
7	beans garbanzo	1		0.89		(TRADER JOE'S	06-28-2014
8	spruuted ca style	1		2.99		(TRADER JOE'S	06-28-2014
9	aavocadus hass bag act	1		3.93		(TRADER JOE'S	06-28-2014
10	aapple bag jazz	2	lb	2.59		(TRADER JOE'S	06-28-2014
11	grocery non taxable	1		0.98		(TRADER JOE'S	06-28-2014
12	creamy salted peanut butler	1		2.49		(TRADER JOE'S	06-28-2014
13	whe wht pita bread	1		4.69		(TRADER JOE'S	06-28-2014
14	grocery non taxable	1		1.38		(TRADER JOE'S	06-28-2014
15		1		0.69	2	(TRADER JOE'S	06-28-2014
16	Total			40.56		(TRADER JOE'S	06-28-2014

Framework : Frontend Overview

Frontend (Streamlit):

User Interface Workflow

1. Allows users to upload files (images or PDFs) via a file uploader component.

2. Displays previews of uploaded images (**Pillow** library)

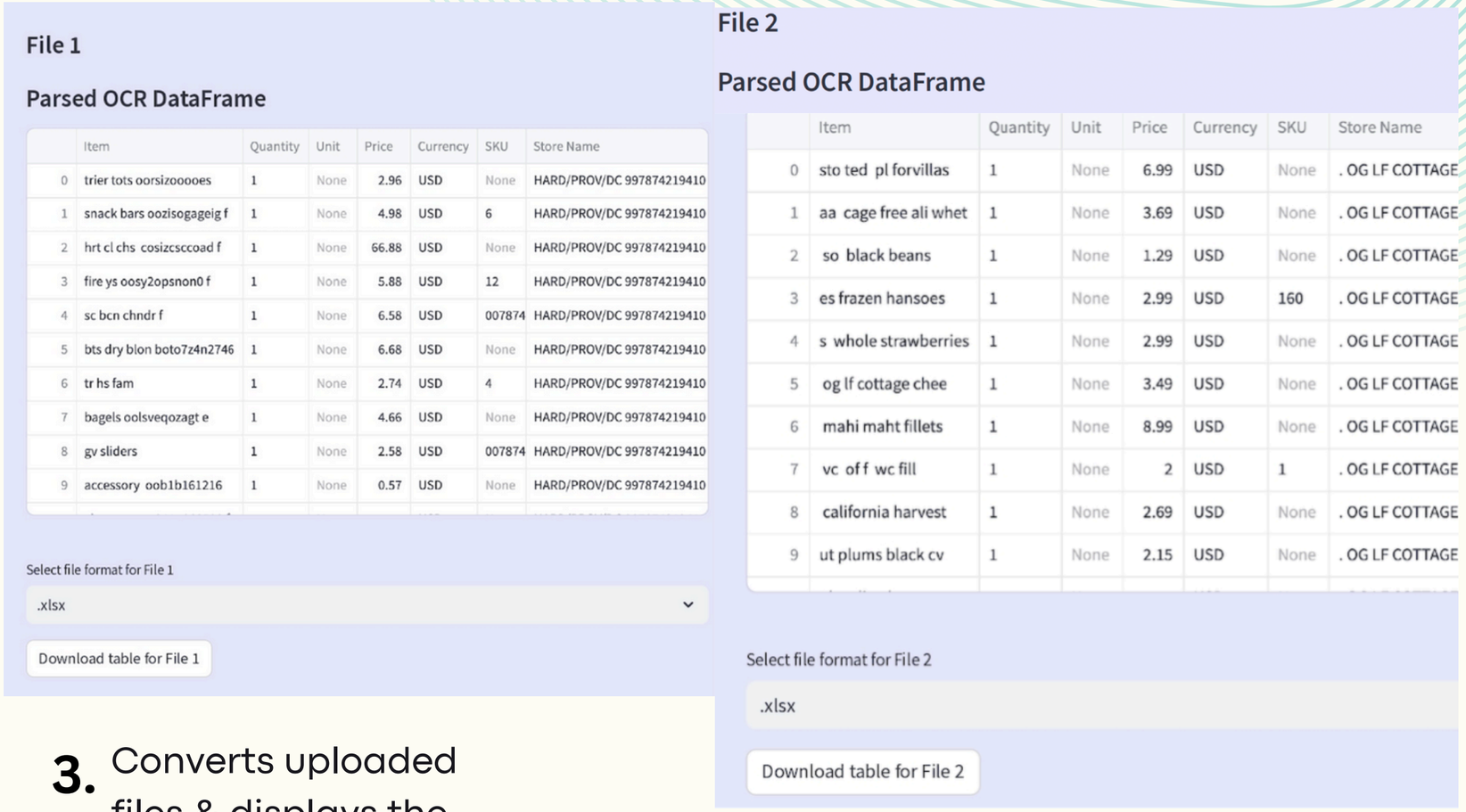
Framework : Frontend Overview

Frontend (Streamlit):

User Interface Workflow

Upon clicking "Convert Files":

- Files are sent to the backend via an HTTP POST request.
- Backend processes the files and returns OCR data.
- Converted data is displayed for each uploaded file separately.
- Users can download the data in their preferred format (Excel or PDF).



3. Converts uploaded files & displays the OCR results.

Conclusion

Challenges

- Optimizing -
  - Image Processing to grayscale/binary.(OpenCV)
  - Choosing correct parameters for the image to string algorithm(tesseract)
- Text processing and cleaning

Key Achievements

- Successful Integration of OCR
- A powerful system capable of converting unstructured data from receipts, invoices, and other documents into structured, actionable information

Next Steps

- Improve Processing Speed
- Integrate with more languages
- Improve Parsing

Demo

Questions?