

OCR Tool for Parsing Invoices and Receipts



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

Computer Vision

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)**.

OCR Receipts Text Detection - retail dataset

Data Card Code (7) Discussion (0) Suggestions (0)

images (20 files)

41 New Notebook Download (55 MB) :

Data Explorer
Version 1 (56.33 MB)

- boxes
- images
 - 0.jpg
 - 1.jpg
 - 10.jpg
 - 11.jpg
 - 12.jpg
 - 13.jpg
 - 14.jpg
 - 15.jpg
 - 16.jpg
 - 17.jpg
 - 18.jpg
 - 19.jpg
 - 2.jpg
 - 3.jpg
 - 4.jpg
 - 5.jpg
 - 6.JPG
 - 7.jpg
 - 8.jpg
 - 9.jpg
- annotations.xml
- receipts.csv

Summary

- 42 files

The screenshot displays a grid of 20 grocery store receipts, each with a thumbnail, file name, and file size. The receipts are from various stores like Whole Foods, Walmart, and Lippo Mall Kemang. The 'Data Explorer' on the right shows the dataset structure with files like boxes, images, annotations.xml, and receipts.csv. A summary at the bottom indicates 42 files.

Framework: Backend Overview

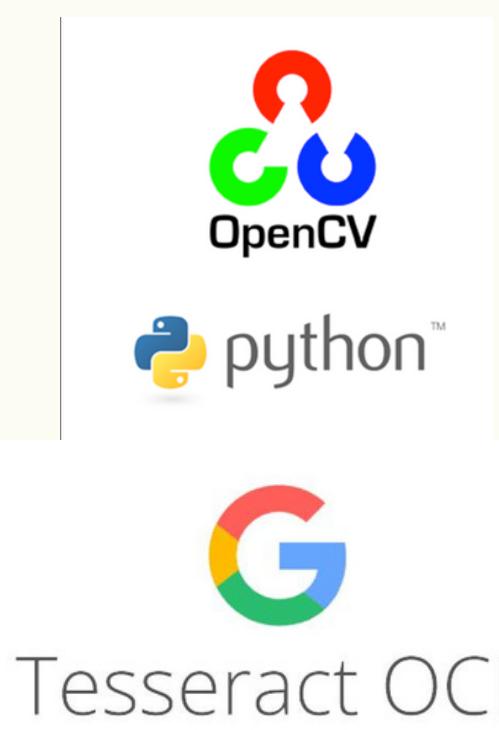
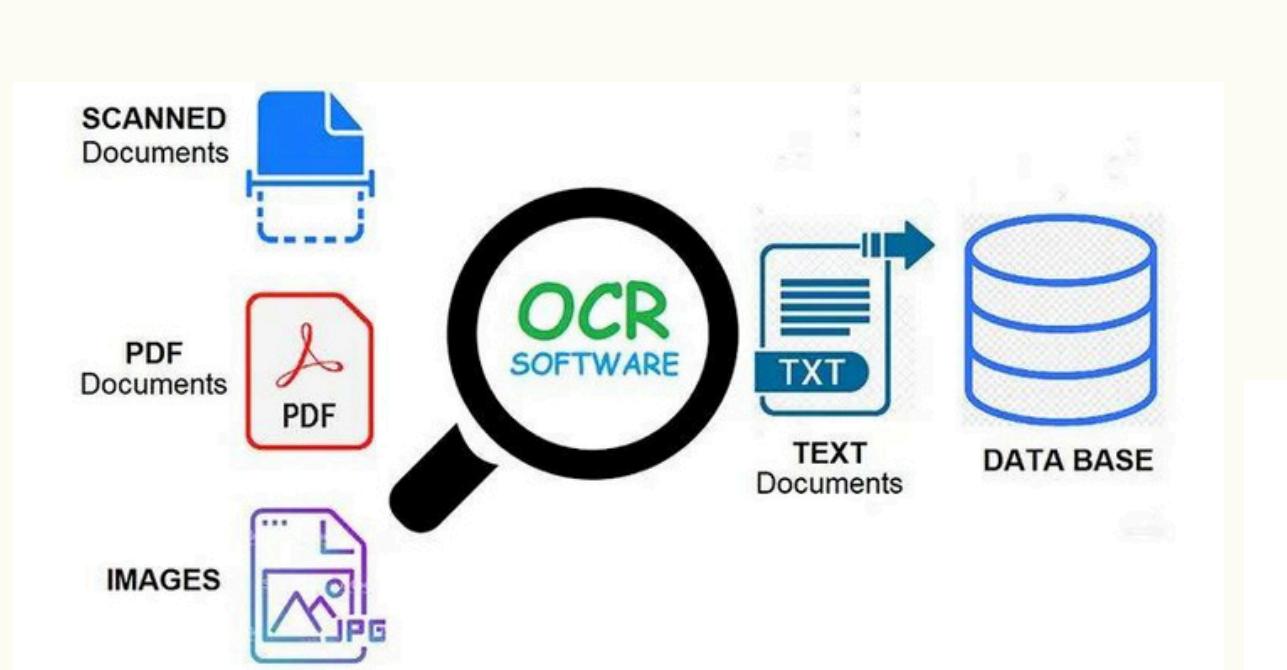


Flask

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...)



Tesseract OCR

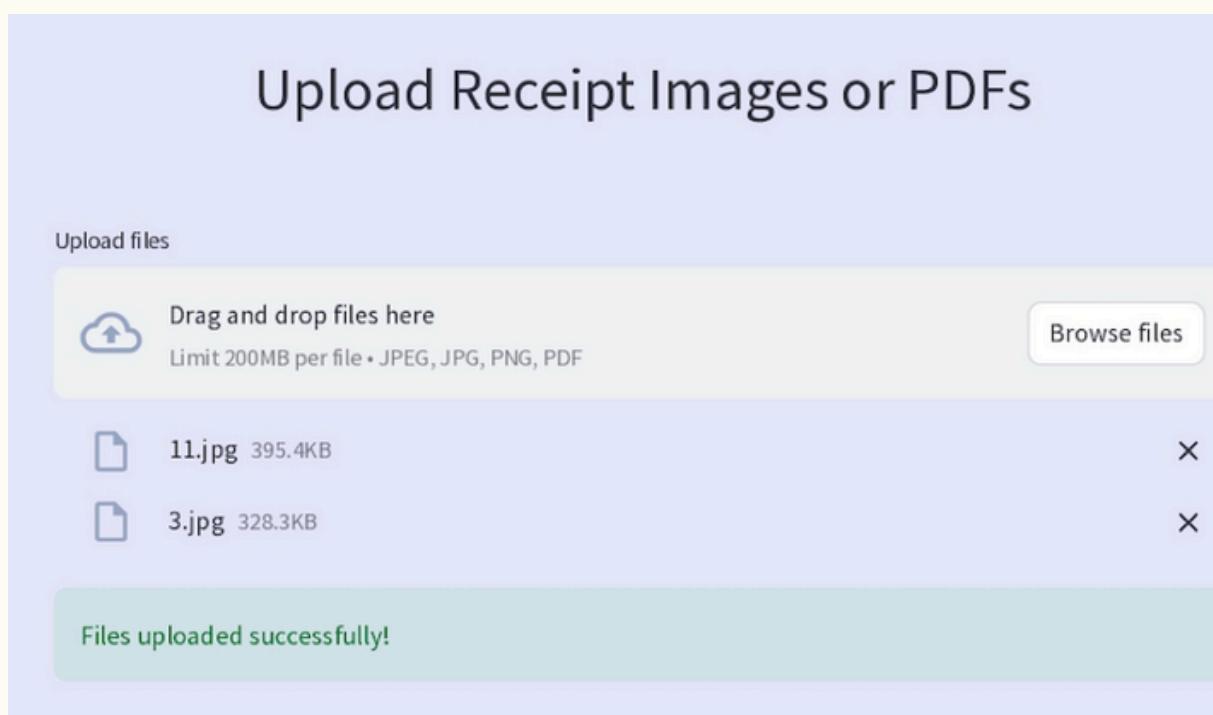
2024-06-22T19-07_export

	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	sprouted 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).

The screenshot shows a Streamlit application interface with two tables side-by-side. Both tables have a light purple header and a white body. The left table is titled 'File 1' and the right table is titled 'File 2'. Each table has a caption 'Parsed OCR DataFrame' above it. Below the tables are dropdown menus for selecting file formats and download buttons.

	Item	Quantity	Unit	Price	Currency	SKU	Store Name
0	trier tots oorsizooooes	1	None	2.96	USD	None	HARD/PROV/DC 997874219410
1	snack bars oozisogageig f	1	None	4.98	USD	6	HARD/PROV/DC 997874219410
2	hrt cl chs cosizcscoad f	1	None	66.88	USD	None	HARD/PROV/DC 997874219410
3	fire ys oosy2opsnon0 f	1	None	5.88	USD	12	HARD/PROV/DC 997874219410
4	sc bcn chndr f	1	None	6.58	USD	007874	HARD/PROV/DC 997874219410
5	bts dry blon boto7z4n2746	1	None	6.68	USD	None	HARD/PROV/DC 997874219410
6	tr hs fam	1	None	2.74	USD	4	HARD/PROV/DC 997874219410
7	bagels oolsveqozagte	1	None	4.66	USD	None	HARD/PROV/DC 997874219410
8	gv sliders	1	None	2.58	USD	007874	HARD/PROV/DC 997874219410
9	accessory oob1b161216	1	None	0.57	USD	None	HARD/PROV/DC 997874219410

	Item	Quantity	Unit	Price	Currency	SKU	Store Name
0	sto ted pl forvillas	1	None	6.99	USD	None	. OG LF COTTAGE
1	aa cage free ali whet	1	None	3.69	USD	None	. OG LF COTTAGE
2	so black beans	1	None	1.29	USD	None	. OG LF COTTAGE
3	es frazen hansoes	1	None	2.99	USD	160	. OG LF COTTAGE
4	s whole strawberries	1	None	2.99	USD	None	. OG LF COTTAGE
5	og lf cottage chee	1	None	3.49	USD	None	. OG LF COTTAGE
6	mahi maht fillets	1	None	8.99	USD	None	. OG LF COTTAGE
7	vc off wc fill	1	None	2	USD	1	. OG LF COTTAGE
8	california harvest	1	None	2.69	USD	None	. OG LF COTTAGE
9	ut plums black cv	1	None	2.15	USD	None	. OG LF COTTAGE

Select file format for File 1
.xlsx

Download table for File 1

Select file format for File 2
.xlsx

Download table for File 2

3. Converts uploaded files & displays the OCR results.

Conclusion

Challenges

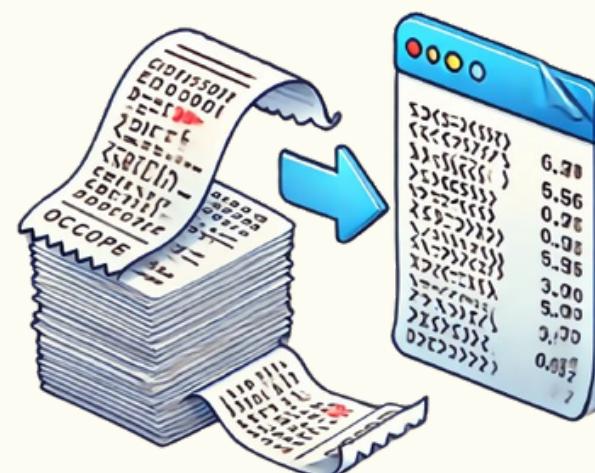
- Optimizing -
 - Image Processing to grayscale/binary.(OpenCV)
 - Choosing correct parameters for the image to string algorithm(tesseract)
- Text processing and cleaning

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?