Мастерская:
CV OCR для справок

Команда #2.6

Задача:

Используя исходный датасет - 16 фотографий справок доноров, создать инструмент для распознавания табличного текста на таких же справках. Фотография может быть любого качества.

Дата	Вид дон-ва	Кол-во	Дата	Вид дон-ва	Кол-во	Дата	Вид дон-ва	Кол-во
1	2	3	4	5	6	7	8	9
14.02.2006	кр/д (бв)	420	15.07.2016	кр/д (бв)	450	04.08.2018	кр/д (бв)	450
11.06.2014	кр/д (бв)	350	11.10.2016	кр/д (бв)	450	26.12.2018	кр/д (бв)	450
30.10.2014	кр/д (бв)	450	21.12.2016	кр/д (бв)	450	29.03.2019	кр/д (бв)	450
13.08.2015	кр/д (бв)	450	21.06.2017	кр/д (бв)	450	11.10.2022	кр/д (бв)	450
30.10.2015	кр/д (бв)	450						

Медрегистратор _____ (подпись)

Отметки о взятии крови (плазмы)

Дата	Вид дон-ва	Кол-во	Дата	Вид дон-ва	Кол-во	Дата	Вид дон-ва	Кол-во
1	2	3	4	5	6	7	8	9
16.06.22	п/ф(бв)	600	06.08.22	п/ф(бв)	610	23.09.22	п/ф(бв)	600
03.10.22	кр/д(бв)	450	10.11.22	п/ф(бв)	600	12.12.22	п/ф(бв)	600
09.01.23	п/ф(бв)	610	31.01.23	п/ф(бв)	610	27.03.23	п/ф(бв)	600
11.04.23	п/ф(бв)	600						

Подходы к решению

#1 Поиск и сегментация таблицы с последующим распознаванием по отдельным ячейкам (автор – Fedor Konovalenko)

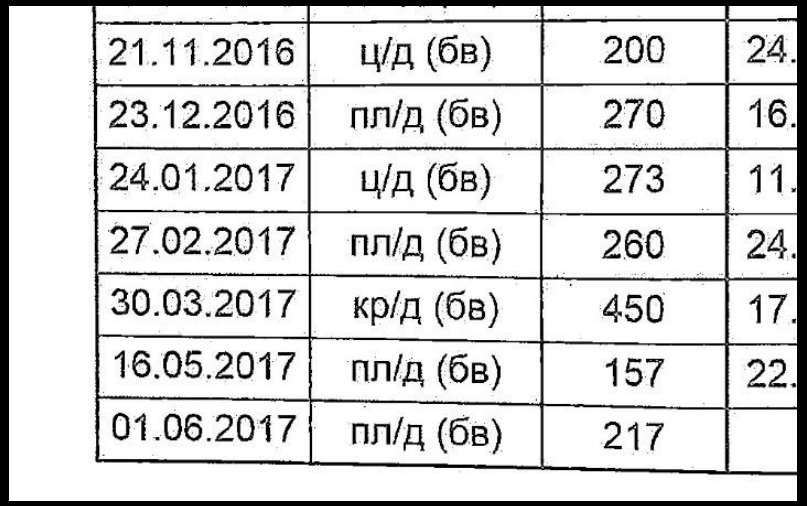
#2 Распознавание изображения целиком с последующим отсевом артефактов, оставшихся от таблицы (автор – Mikhail Maresin)

#1 Распознавание по ячейкам

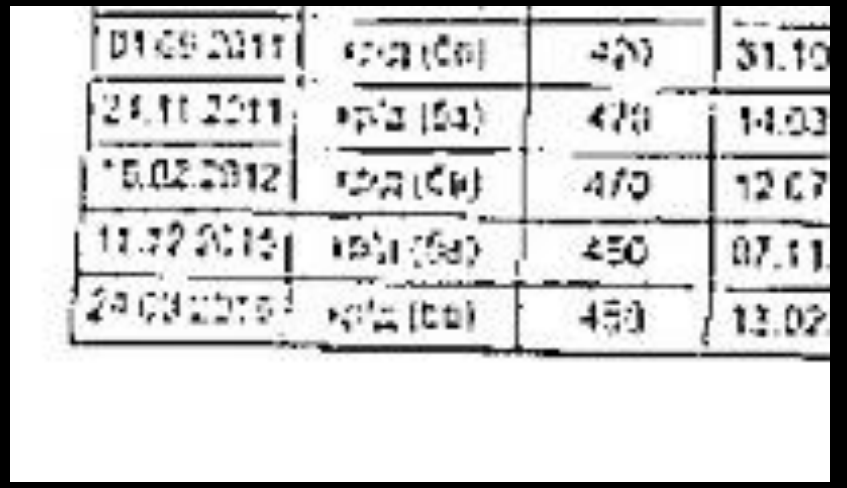
- Препроцессинг (openCV)
- Поиск таблицы (img2table) и координат ее ячеек
- Проход по ячейкам с распознаванием строк (tesseract) в режиме распознавания цифр
- Проход по ячейкам с распознаванием строк (tesseract) в режиме распознавания букв
- Запись найденного в массив и очистка
- Сортировка и сборка в датафрейм (регулярные выражения)

Препроцессинг

- Обрезка верхней части –> оттенки серого –> размытие для удаления шумов –> максимизация контраста
- **Проблема:** изображения с разным разрешением требуют применения разных масок препроцессинга. Сложно улучшить хорошее изображение а вот плохое очень легко сделать еще хуже.

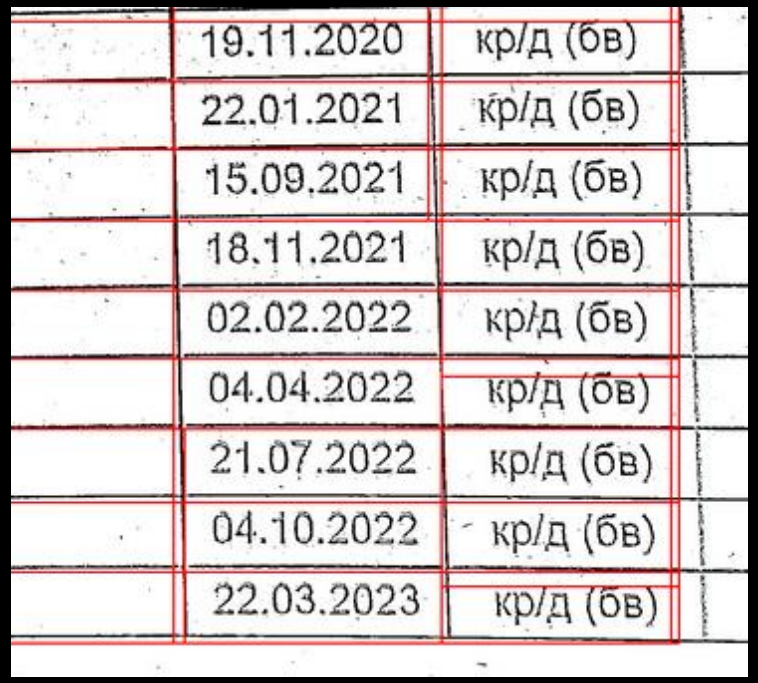


21.11.2016	ц/д (бв)	200	24.
23.12.2016	пл/д (бв)	270	16.
24.01.2017	ц/д (бв)	273	11.
27.02.2017	пл/д (бв)	260	24.
30.03.2017	кр/д (бв)	450	17.
16.05.2017	пл/д (бв)	157	22.
01.06.2017	пл/д (бв)	217	

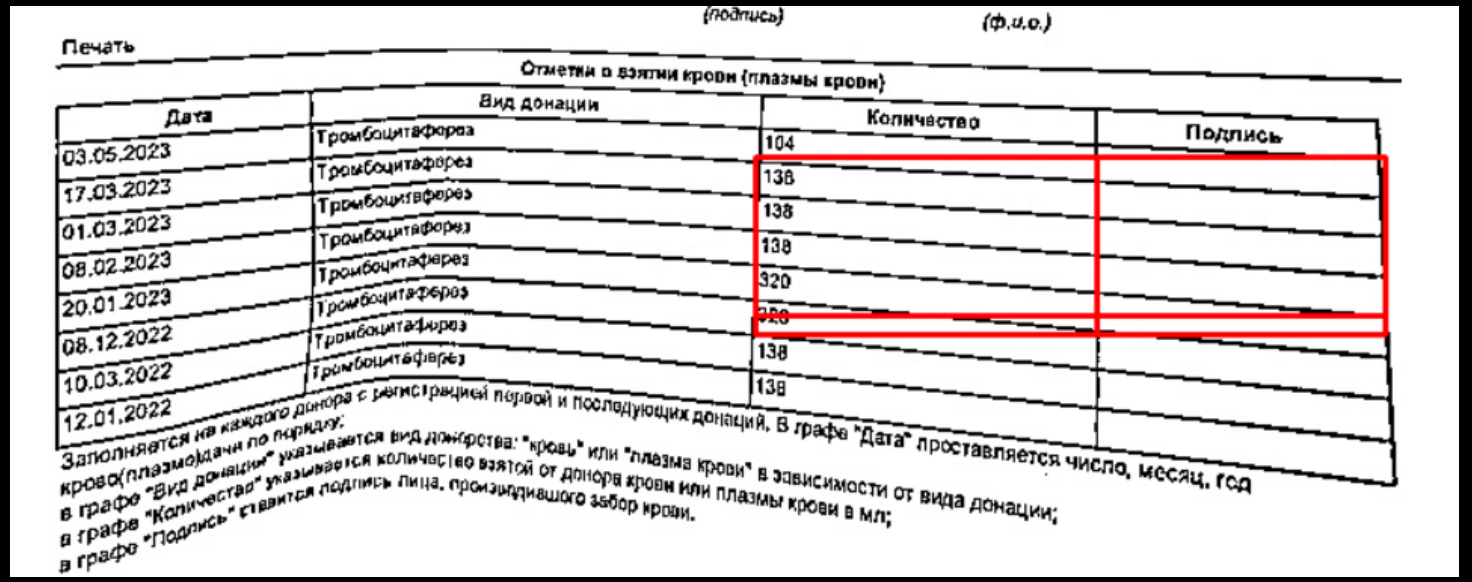


Поиск координат ячеек

- Проблема №1 – дубли. Решается обработкой полученного датасета
- Проблема №2 – мятые бумажки



	19.11.2020	кр/д (бв)
	22.01.2021	кр/д (бв)
	15.09.2021	кр/д (бв)
	18.11.2021	кр/д (бв)
	02.02.2022	кр/д (бв)
	04.04.2022	кр/д (бв)
	21.07.2022	кр/д (бв)
	04.10.2022	кр/д (бв)
	22.03.2023	кр/д (бв)



(подпись) (ф.и.о.)

Печать

Отметки о взятии крови (плазмы крови)

Дата	Вид донации	Количество	Подпись
03.05.2023	Тромбоцитафереа	104	
17.03.2023	Тромбоцитафереа	138	
01.03.2023	Тромбоцитафереа	138	
08.02.2023	Тромбоцитафереа	138	
20.01.2023	Тромбоцитафереа	920	
08.12.2022	Тромбоцитафереа	920	
10.03.2022	Тромбоцитафереа	138	
12.01.2022	Тромбоцитафереа	138	

Заполняется на каждого донора с регистрацией первой и последующих донаций. В графе "Дата" проставляется число, месяц, год кровo(плазмo)дачи по порядку;
в графе "Вид донации" указывается вид донорства: "кровь" или "плазма крови" в зависимости от вида донации;
в графе "Количество" указывается количество взятой от донора крови или плазмы крови в мл;
в графе "Подпись" ставится подпись лица, производившего забор крови.

Сбор данных

Дата	Вид дон-ва	Кол-во	Дата	Вид дон-ва	Кол-во	Дата	Вид дон-ва	Кол-во
1	2	3	4	5	6	7	8	9
22.08.19	к/д(бв)	450	22.10.19	т/ф(бв)	525	21.11.19	т/ф(бв)	442
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	т/ф(бв)	345
25.08.20	т/ф(бв)	435	11.03.21	п/ф(бв)	600	07.04.21	п/ф(бв)	600
22.04.21	п/ф(бв)	600	31.08.21	т/ф(бв)	356	22.09.21	т/ф(бв)	345
29.12.21	т/ф(бв)	500	08.02.22	т/ф(бв)	342	23.03.22	т/ф(бв)	342
01.06.22	т/ф(бв)	342	12.07.22	т/ф(бв)	430	08.08.22	т/ф(бв)	530
25.10.22	т/ф(бв)	345	05.12.22	т/ф(бв)	368	. .		

Проход слева направо и сверху вниз

Используется структура таблицы: “Дата – вид донорства + класс крови - объем”

Очистка и сборка датафрейма

Сырой список

o\n', '51\n', '20.04.2011\n', '6\n', '350\n', 'т\n', '6\n',
'|\n', '450\n', 'B\n', 'ранние\n', '13.03.2015\n', '68\n',
'450\n', '02.08.2017\n',

После очистки и регулярных выражений

'20.04.2011', '6', '350', 'т', '6', 'unknown', '450', 'B',
'unknown', '13.03.2015'

Очистка и сборка датафрейма

«Главное – распознать дату, а дальше – да поможет нам Левенштейн»

```
pd.read_csv('result/recognized/236000.csv')
```

	Unnamed: 0	Дата донации	Тип донации	Класс крови
0	0	2020-11-25	Безвозмездно	Цельная кровь
1	1	2021-02-26	Безвозмездно	Цельная кровь
2	2	2021-11-09	Безвозмездно	Цельная кровь
3	3	2022-09-16	Безвозмездно	Цельная кровь
4	4	2023-02-17	Безвозмездно	Цельная кровь

Распознали все

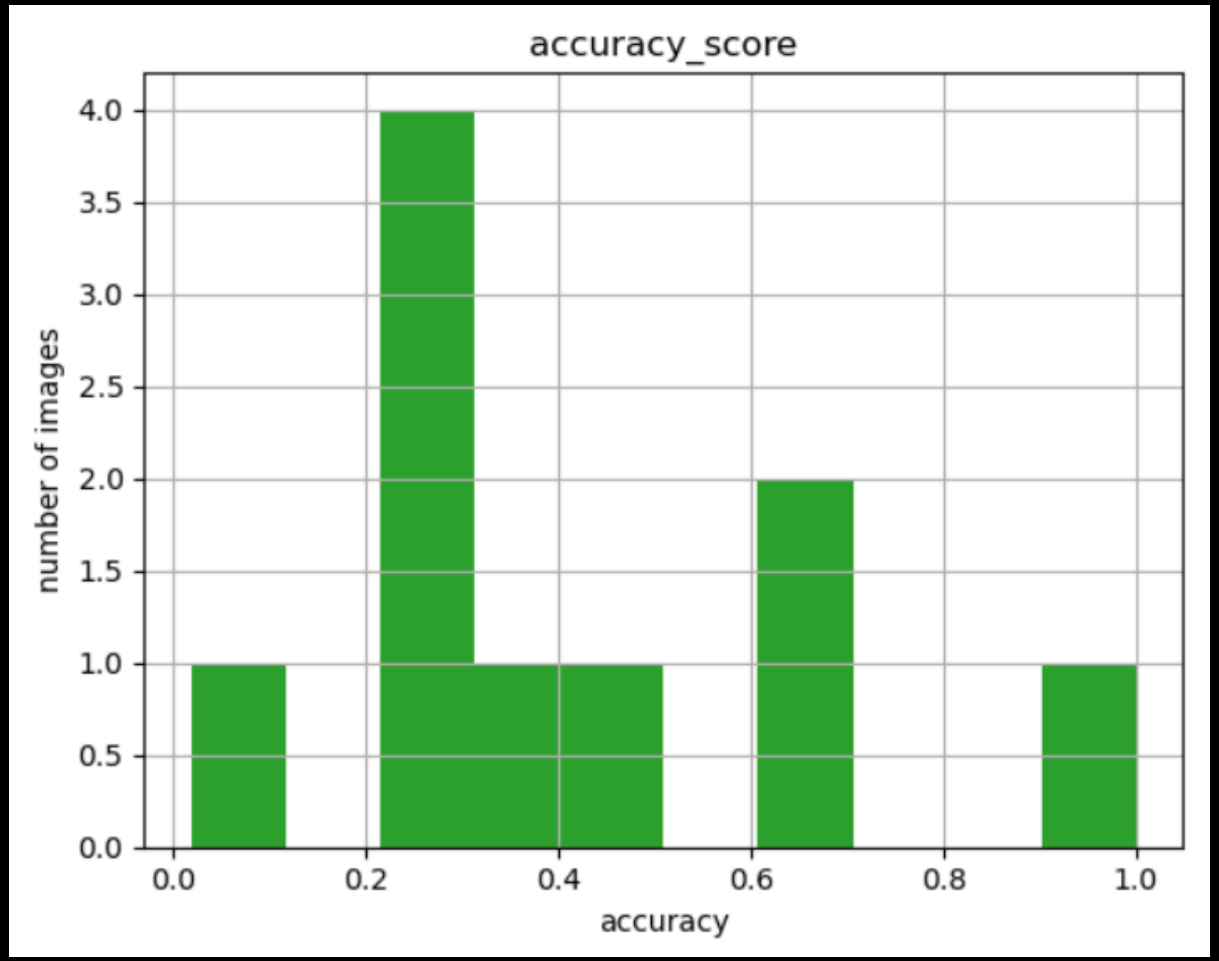
	Unnamed: 0	Дата донации	Тип донации	Класс крови
0	0	2009-01-14	unknown	unknown
1	1	2009-07-14	Безвозмездно	Цельная кровь
2	2	2010-01-25	unknown	unknown
3	3	2011-02-07	Безвозмездно	Цельная кровь
4	4	2011-08-08	unknown	unknown
5	5	2012-02-29	unknown	unknown
6	6	2016-11-21	unknown	unknown
7	7	2017-01-24	unknown	unknown
8	8	2017-02-27	Безвозмездно	Цельная кровь

Распознали не все

#1 Результат

Распознать хоть что-то получилось на 10 изображениях из 16

Среднее значение accuracy – 0,44



Проблемы и слабые места

- Сделана «распознавалка справок формы 405»
- В исходных данных донации только бесплатные
- ‘кр/д (бв) ’ – как много в этом звуке…

#2 Распознавание целиком

- Препроцессинг (обрезка верхней части)
- Распознавание «как есть» (Easyocr)
- Очистка от пространственных меток
- Кластеризация вокруг центров блоков с текстом
- Сборка таблицы

#2 Результат

- Точнее, чем способ #1
- Более требователен к вычислительным ресурсам:

	Name	Tag	Status	Created	Size	Ac
☐	easyocr d3deaead46e0	latest	Unused	2 minutes ago	8.55 GB	▶
☐	tesseract a34cdc98109b	latest	Unused	13 hours ago	1.98 GB	▶

	0	1	2	3	4	5
0	14.04.2014	кр/д (бв)	450	21.04.2020	кр/д (бв)	450
1	30.05.2018	кр/д (бв)	450	27.07.2020	кр/д (бв)	350
2	30.07.2018	кр/д (бв)	450	19.11.2020	кр/д (бв)	450
3	05.10.2018	кр/д (бв)	450	22.01.2021	кр/д (бв)	450
4	05.12.2018	кр/д (бв)	450	15.09.2021	кр/д (бв)	450
5	06.02.2019	кр/д (бв)	450	18.11.2021	кр/д (бв)	450
6	11.04.2019	кр/д (бв)	450	02.02.2022	кр/д (бв)	450
7	05.07.2019	кр/д (бв)	450	04.04.2022	кр/д (бв)	450
8	10.09.2019	кр/д (бв)	450	21.07.2022	кр/д (бв)	450
9	22.11.2019	кр/д (бв)	450	04.10.2022	кр/д (бв)	450
10	24.01.2020	кр/д (бв)	450	22.03.2023	кр/д (бв)	450

Промежуточный итог

Вариант #1

Вариант #2

Микросервис (вариант #1)

- Загрузка изображения
- Выдача результата распознавания в JSON
- Обработка исключений – статусы:

“изображение распознано”,

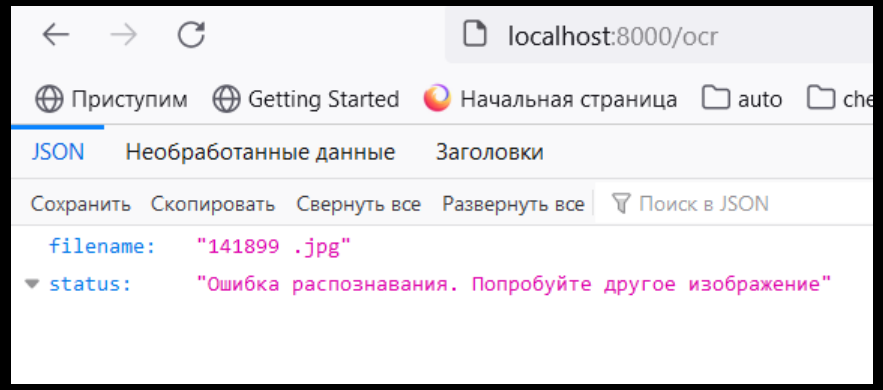
“изображение распознано не полностью”,

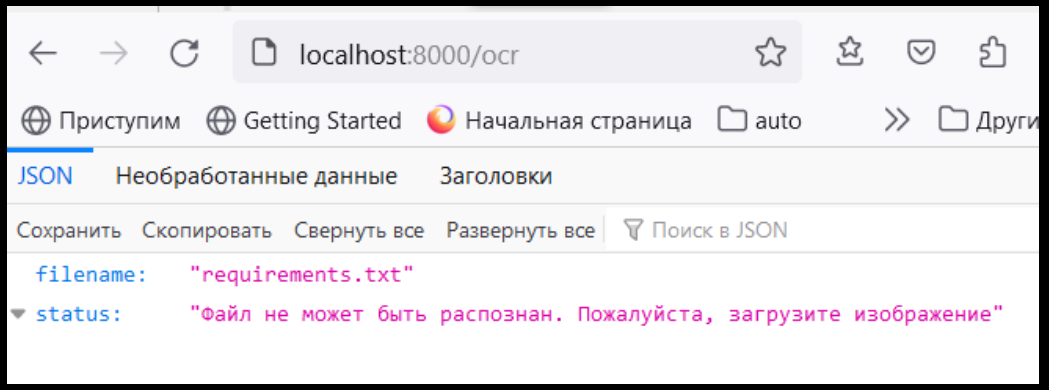
“изображение не распознано”,

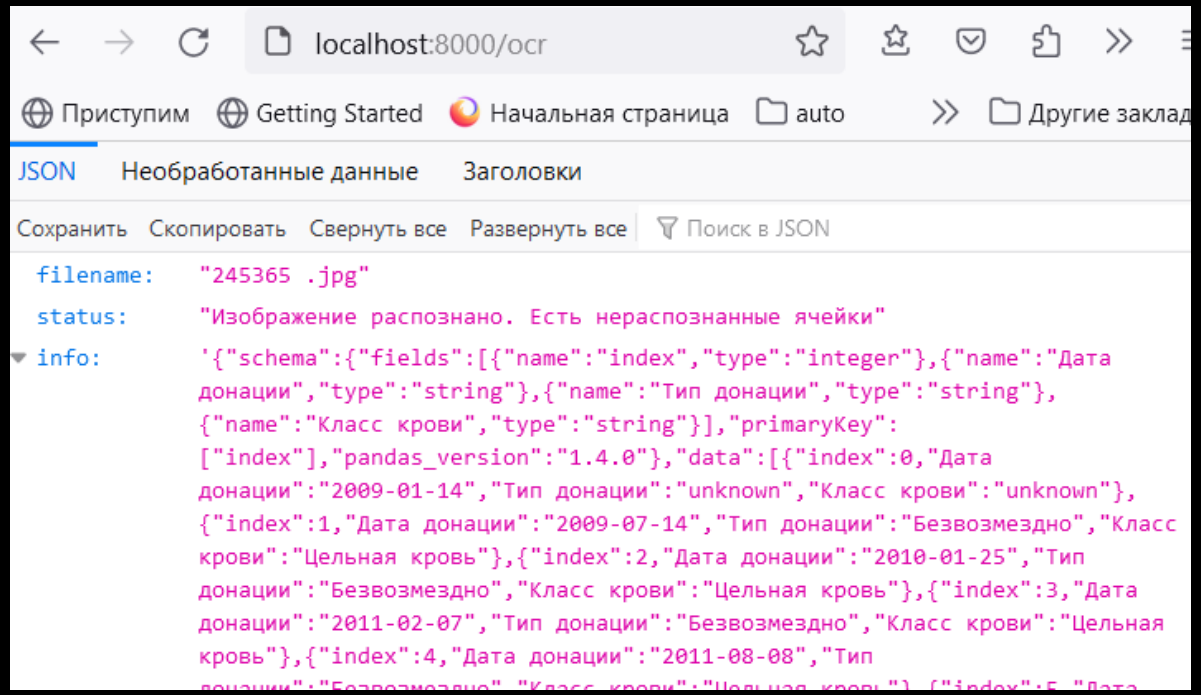
“загруженный файл не является изображением”*

* не представим в виде numpy.array

Микросервис (вариант #1)







¯_(ツ)_/¯

- Образ собирается
- Контейнер запускается
- Не работает :(

Наиболее вероятная причина – img2table при поиске таблицы выдает разный результат при запуске вне контейнера и в контейнере

«Сегодня я многое понял»

Если какая-то библиотека тебе понравилась – проверь, может быть, она понравилась **только** тебе