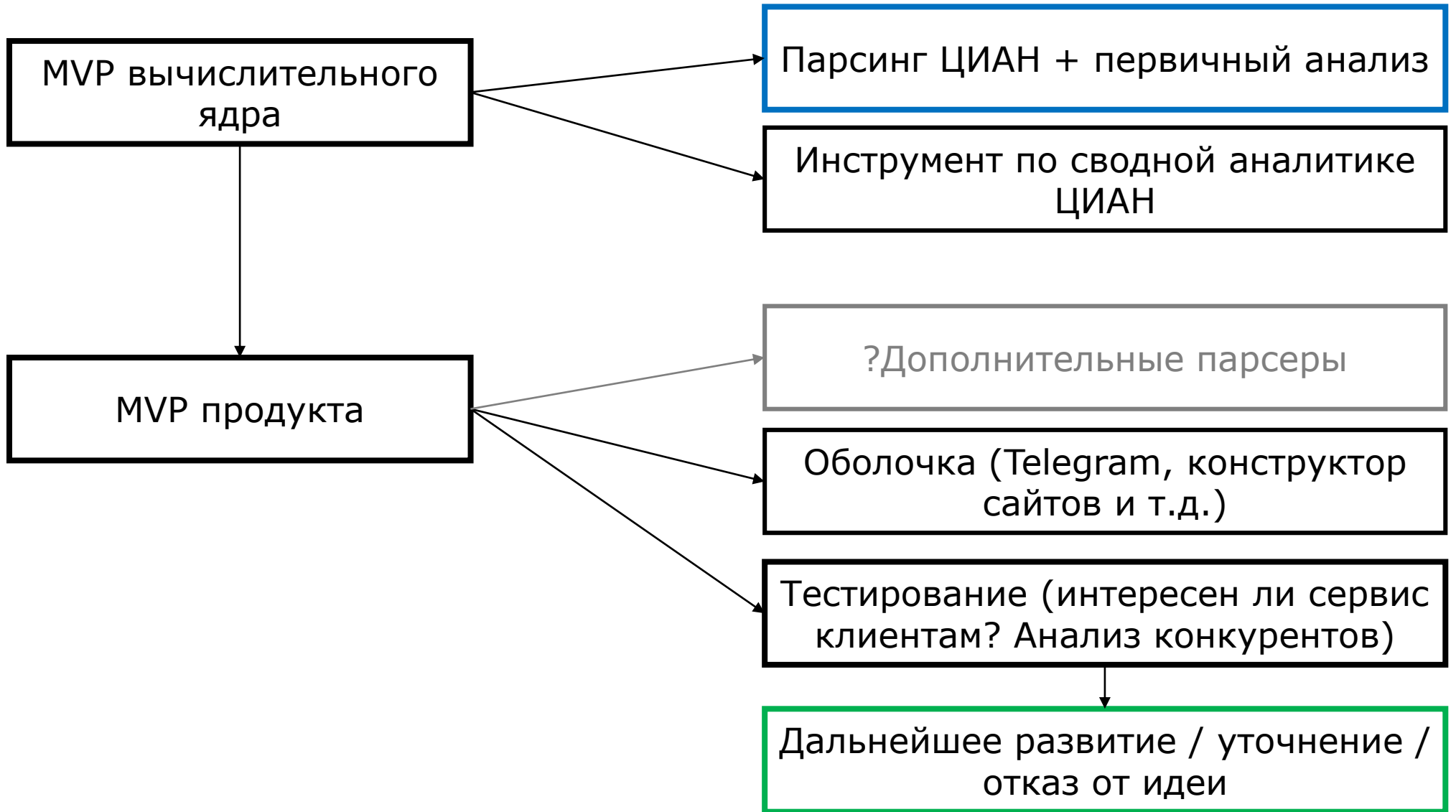


# Схема проекта

Сервис поиска интересных предложений для покупки недвижимости



# Спринт №1: Парсинг ЦИАН + первичный анализ

**Цели:** выгрузить объявления по продаже жилой недвижимости третьего транспортного кольца Москвы, стоимость которых ниже рыночной оценки.

**Входные данные:** данные о продаже недвижимости в Москве с сайта ЦИАН.

**Проблемы:** наличие фейков.

## **Задачи:**

1. Разработать парсер датасета с сайта ЦИАН (минимальный набор факторов: широта и долгота расположения дома, общая площадь, количество комнат, материал стен, этаж квартиры и этажность здания, адрес, округ, тип продажи);
2. Провести обработку пропусков и поиск аномалий;
3. Первичный анализ факторов, фича-инжиниринг исходя из бизнес логики (пример – расстояние до центра);
4. Провести вывод объявлений в .xlsx документ со стоимостью ниже выборочного среднего значения похожих объявлений.

**Методы и библиотеки:** Jupyter Notebook, Python 3, pandas, scikit-learn, matplotlib, geopy.

# Парсинг ЦИАН: обзор открытой библиотеки cianparser

**URL:** <https://github.com/lenarsaitov/cianparser>

## Возможности:

- Конфигурация входных данных по 7 параметрам: аренда/продажа, город, вид жилья – квартира/таунхаус/и т.д., кол-во комнат, страница начала сбора данных, страница конца сбора данных, скорость сбора, собственник;
- Конфигурация выходных данных по 2 параметрам: кодировка, выгрузка в формате .csv.

## Выгружаемые данные:

Признаки, получаемые в ходе сбора данных

- **district** - район
- **underground** - метро
- **street** - улица
- **floor** - этаж
- **floors\_count** - общее количество этажей
- **total\_meters** - общая площадь
- **living\_meters** - жилая площади
- **kitchen\_meters** - площадь кухни

- **rooms\_count** - количество комнат
- **year\_construction** - год постройки здания
- **price** - стоимость
- **price\_per\_m2** - стоимость на квадратный метр
- **author** - автор объявления
- **author\_type** - тип автора
- **phone** - номер телефона в объявлении
- **link** - ссылка на объявление

Возможные значения поля **author\_type**: **real\_estate\_agent** - агентство недвижимости, **homeowner** - собственник, **realtor** - риелтор, **official\_representative** - ук оф.представитель, **representative\_developer** - представитель застройщика, **developer** - застройщик, **unknown** - без указанного типа.

# Парсинг ЦИАН: обзор открытой библиотеки cianparser

## Тестирование:

[https://cian.ru/cat.php?engine\\_version=2&p=1&region=1&offer\\_type=flat&deal\\_type=rent&room2=1&room3=1&with\\_neigh\\_bors=0&type=4](https://cian.ru/cat.php?engine_version=2&p=1&region=1&offer_type=flat&deal_type=rent&room2=1&room3=1&with_neigh_bors=0&type=4) (14.04.23)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	author	author_ty	link	city	deal_type	accommo	floor	floors_co	rooms_co	total_me	price_per	price_per	commissi	year_of_c	living_me	kitchen_n	phone	district	street	underground		
2	Excellent real estate	real_esta	https://w	Москва	rent	flat	11	15	2	45	1444	65000	30	2019	30	10	7.96E+10	Даниловс	Автозаво	ЗИЛ		
3	ЦИАН	real_esta	https://w	Москва	rent	flat	4	9	2	48	875	42000	50	1963	35	6	7.97E+10	Марфино	Академи	Фонвизинская		
4	ЦИАН	real_esta	https://w	Москва	rent	flat	11	15	2	41	1585	65000	50	2019	24	12	7.96E+10	Даниловс	Автозаво	ЗИЛ		
5	Excellent real estate	real_esta	https://w	Москва	rent	flat	4	9	2	45	933	42000	30	1963	30	10	7.97E+10	Марфино	Академи	Фонвизинская		
5	ID 18835544	official_r	https://w	Москва	rent	flat	10	12	3	121	7024	850000	0	2015	-1	-1	7.96E+10	Тверской	Охотный ф	Охотный ряд		
7	ID 47020496	homeowr	https://w	Москва	rent	flat	26	44	2	50	2140	107000	0	2018	33.3	-1	7.5E+10	Пресненс	Ходынская	Улица 1905 года		
8	ID 18835544	official_r	https://w	Москва	rent	flat	5	12	2	80	6875	550000	0	2015	-1	-1	7.96E+10	Тверской	Охотный ф	Охотный ряд		
9	ID 24309666	homeowr	https://w	Москва	rent	flat	9	16	2	40	2125	85000	0	2018	22	9	7.91E+10	Филевски	Большая	Фили		
0	ID 18835544	official_r	https://w	Москва	rent	flat	7	12	2	112	8928	1000000	0	2015	-1	-1	7.96E+10	Тверской	Охотный ф	Охотный ряд		
1	ID 18835544	official_r	https://w	Москва	rent	flat	10	12	2	88	9090	800000	0	2015	-1	-1	7.96E+10	Тверской	Охотный ф	Охотный ряд		
2	Лиел Инвестментс Л	official_r	https://w	Москва	rent	flat	42	75	2	98	3469	340000	0	2013	-1	-1	7.92E+10	Пресненс	1-й Красн	Деловой центр		
3	Лиел Инвестментс Л	official_r	https://w	Москва	rent	flat	52	75	2	100	3300	330000	0	2013	-1	-1	7.92E+10	Пресненс	1-й Красн	Деловой центр		
4	Лиел Инвестментс Л	official_r	https://w	Москва	rent	flat	52	75	2	101	3168	320000	0	2013	-1	-1	7.92E+10	Пресненс	1-й Красн	Деловой центр		
5	Лиел Инвестментс Л	official_r	https://w	Москва	rent	flat	56	75	2	90	3666	330000	0	2013	-1	-1	7.92E+10	Пресненс	1-й Красн	Деловой центр		
6	ID 32373776	homeowr	https://w	Москва	rent	flat	2	9	2	56	803	45000	0	1973	46	8	7.5E+10	Ивановск	Молостое	Новогиреево		
7	ЦИАН	real_esta	https://w	Москва	rent	flat	3	10	2	75	1733	130000	0	2020	55	10	7.97E+10	Пресненс	Мантулин	Выставочная		

**Ограничения:** скорость (возможна блокировка IP), объем (28\*54=1512 объявлений, сайт выдает списки максимум до 54 страницы поискового запроса).

**Требования:** Python 3 (версия 3.8 и выше), установка библиотеки transliterate.

**Дополнительно:** актуальная библиотека - последнее обновление 2 месяца назад.

# Парсинг ЦИАН: фильтрация входных данных коммерческой недвижимости

**Задача:** разработать на базе `ciaparser` парсер для фильтров алгоритма экспертной оценки коммерческой недвижимости.

## Обзор кода:

*Структура парсера:*

Шаг 1. Загрузка HTML/CSS кода страницы списка объявлений;

Шаг 2. Выделение кусков кода с требуемой информацией (автор, стоимость и т.д.) по ключевым словам.

HTML код страницы грузится, но его структура для коммерческой недвижимости отличается от обычной аренды, поэтому нужная информация находится под другими названиями элементов.

Структура кода самого парсера перегружена ненужными для нас элементами (около 920 строк кода, нет комментариев).

**Вывод:** модифицировать дольше, чем написать свой парсер с нуля.

# Парсинг ЦИАН: фильтрация входных данных коммерческой недвижимости

```
1 #Коммерческая
2 #url = 'https://www.cian.ru/sale/commercial/277193588/'
3 #Краткосрочная аренда
4 url = 'https://www.cian.ru/rent/flat/285274355/'
5
6 #Загрузка всего HTML кода страницы
7 session = cloudscraper.create_scraper()
8 res = session.get(url=url)
9 res.raise_for_status()
10 html = res.text
11
12 #Показать первые 200 символов исходного кода страницы
13 soup = BeautifulSoup(html, 'lxml')
14 print(str(soup.get_text())[0:200])
```

**Краткосрочная аренда  
жилой недвижимости**



Сдам двухкомнатные апартаменты 171м² ул. Охотный Ряд, 2, Москва, ЦАО, р-н Тверской м. Охотный ряд - база ЦИАН, объявление 285274355

**Коммерческая  
недвижимость**



**HTTPError: 403 Client Error: Forbidden for url:**  
8/

**Вывод:** ЦИАН защищает (капча) страницы объявлений от парсинга (в отличие от страниц поискового запроса). Судя по алгоритму cianparser данной проблемы не было 2 месяца назад.

# Парсинг ЦИАН: фильтрация входных данных коммерческой недвижимости

## НО:

(исходный код страницы поискового запроса)

год постройки и материал стен (не отображается на странице поискового запроса)

```
84 "heatingType":null,"classType":null,"accessType":null,"materialType":null,"buildYear":1967,"  
85  
86
```

координаты объекта

```
83  
84 , "boundedBy": {"lowerCorner": {"lng": 38.684018, "lat": 43.760146}, "upperCorner":  
85  
86
```

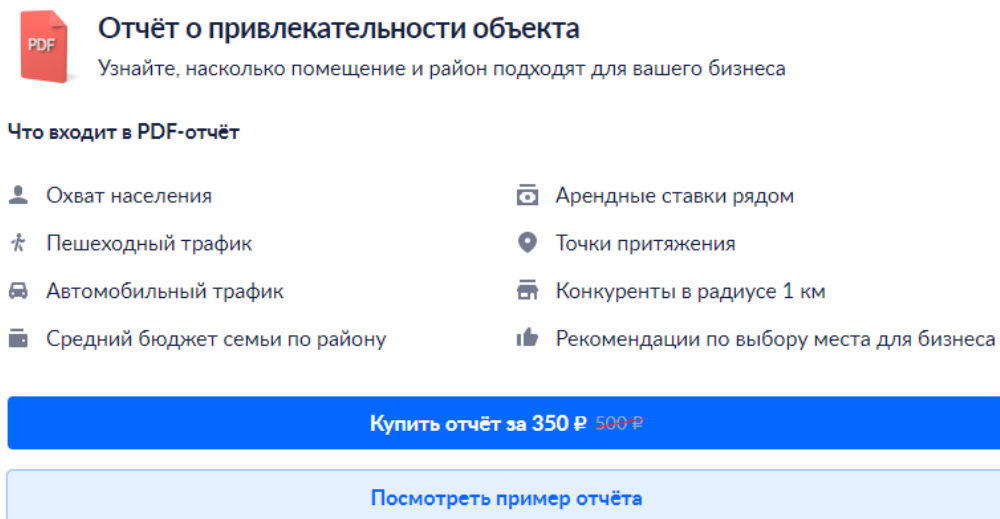
**Вывод:** однако вся информация (факторы) об объявлениях находится на странице поискового запроса.

# Парсинг ЦИАН: фильтрация входных данных коммерческой недвижимости

## ЦИАН против парсинга:

- мгновенный сброс парсера на страницах объявления при использовании популярных библиотек парсинга (капча) – **фактически запрещает отправлять напрямую объявление в разрабатываемое приложение;**
- **нужно регулярно обновлять парсер:** [парсер который работал год назад](#) – не работает, парсер который работал 2 месяца работает частично (ЦИАН проводит регулярное обновление названий переменных в коде).

Возможная причина – ЦИАН сам является нашим конкурентом.



Некоторых элементов PDF-отчета нет даже в примере

[https://files.cian.ru/files/commercial/geo-analytics/report\\_example.pdf](https://files.cian.ru/files/commercial/geo-analytics/report_example.pdf)

**Выводы:** в будущем необходим сервис для парсинга ([например](#)), имеет смысл изучить вопрос легальности парсинга ЦИАН.