

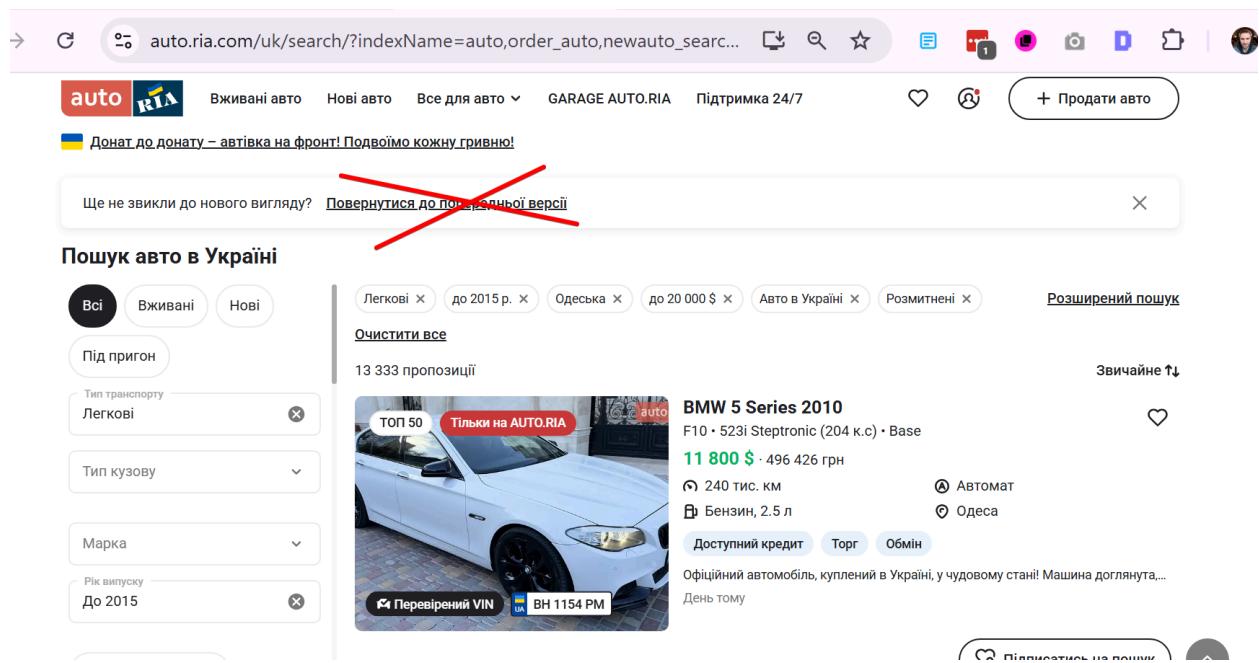
Реализуем парсер Autoria в виде Python скрипта

Входные данные

В качестве входных данных (файл input.txt лежит в корневой директории скрипта) для парсера вы подаете ссылку на выдачу сайта авториа. Например:

[https://auto.ria.com/uk/search/?search_type=1&category=1&all\[0\].any\[0\].state=2&all\[0\].any\[1\].state=8&all\[0\].any\[2\].state=6&all\[0\].any\[3\].state=1&all\[0\].any\[4\].state=16&all\[0\].any\[5\].state=20&all\[0\].any\[6\].state=24&price\[0\]=1&price\[1\]=3000&abroad=0&customs_cleared=1&order=7&republi shed_last=4](https://auto.ria.com/uk/search/?search_type=1&category=1&all[0].any[0].state=2&all[0].any[1].state=8&all[0].any[2].state=6&all[0].any[3].state=1&all[0].any[4].state=16&all[0].any[5].state=20&all[0].any[6].state=24&price[0]=1&price[1]=3000&abroad=0&customs_cleared=1&order=7&republi shed_last=4)

Важный момент - работаем именно с текущей (новой) версие сайта



The screenshot shows the search results for a car query on the auto.ria.com website. A prominent red 'X' is drawn over a message at the top of the page that reads "Ще не звикли до нового вигляду? Поверніться до попередньої версії". The main search results area displays a listing for a "BMW 5 Series 2010" with a price of "11 800 \$". The listing includes details like "F10 • 523i Steptronic (204 к.с.) • Base", "240 тис. км", "Бензин, 2.5 л", and "Одеса". There are also buttons for "Доступний кредит", "Торг", and "Обмін". The left sidebar contains filters for "Bci", "Легкові", "до 2015 р.", "Одеска", "до 20 000 \$", and "Розмітнені". The right sidebar has a "Розширений пошук" section.

Сбор ссылок на товары со страницы поиска

Парсер собирает ссылки на результаты (отдельные объявления) с первой страницы выдачи (каталога) объявлений, далее переходит на следующую страницу и так до конца. Ранее было замечено, что при загрузке страницы не всегда сразу прогружается блок с пагинацией. Соответственно, нужно дождаться его загрузки, чтобы не утерять данные, которые доступны на следующих страницах

auto.ria.com/uk/search/?indexName=auto,order_auto,newauto_searc...

Вживані Нові

рион

автоспорту

їві

узову

а

туску

15

ревірений VIN

Перевірений VIN UA BE 8100 EO

Перевірений VIN UA BH 4071 TM

Продаю додгянутий BMW 5 серії F10 – автомобіль бізнес-класу, я...

19 годин тому

Mercedes-Benz Viano 2013

W639

16 000 \$ · 673 120 грн

444 тис. км

Дизель, 2.14 л

Автомат

Одеса

Доступний кредит Торг

офіційний. оригінальний пасажир. автомобіль не вимагає жодни

4 дні тому

Показувати по:

1 2 3 4 ... 667 >

Сбор данных о каждом товаре

Сбор данных происходит со страницы объявления. С каждого объявления выгружаем: заголовок объявления, телефон (если доступен), имя (если доступно), город, дату публикации, цену, ссылку на объявление. Телефон получаем кликом по плашке с телефоном.

← → ⌂ auto.ria.com/uk/auto_opel_vivaro_39133008.html

auto RIA Вживані авто Нові авто Все для авто GARAGE AUTO.RIA Підтримка 24/7

Донах до донату – автівка на фронт! Подвоїмо кожну гривню!

Ще не звікли до нового вигляду? Повернутися до пошуку

AUTO.RIA.com > Легкові з пробігом > Черкаська область >

Opel Vivaro 2011

Продавець
Міщенко Віталій

(097) 270 28 56

Попросити продавця перетелефонувати

Передзвоніть мені

Оцініть продавця

☆ ☆ ☆ ☆ ☆

Opel Vivaro 2011

I покоління/A (FL)

11 200 \$ · 471 180 грн

Розширені історія авто

Доступна перевірка · від 550 грн

Продавець
Міщенко Віталій

★ Оцінити продавця >

5+ років працює з AUTO.RIA

(097) XXX XX XX



Ще не звикли до нового вигляду? [Повернутися до попередньої версії](#)

AUTO.RIA.com > Легкові з пробігом > Черкаська область > Черкаси > Opel > Vivaro > Opel Vivaro Черкаси



Opel Vivaro 2011

I покоління/A (FL)

11 200 \$ • 471 184 грн

Розширенна історія авто

Доступна перевірка • від 550 грн



Продавець
Міщенко Віталій

→ C auto.ria.com/uk/auto_opel_vivaro_39133008.html

AUTO.RIA

Стан кузова Технічний стан Стан салону Тестова поїздка

від 1 110 грн Замовити

Знаєте більше про це авто?
Напишіть коментар, і ми перевіримо, щоб все було чесно

Розказати правду

Мої досягнення в перевірках > Про безпечні угоди >

Оголошення створене 07.11.2025

Переглядів авто 5

Додано в Обране 0

ID авто 39133008

Оформити розстрочку на Opel Vivaro 2011

Коли треба вкластиς у

Продавець
Міщенко Віталій

Оцінити продавця >
5+ років працює з AUTO.RIA

(097) XXX XX XX

Важный момент - перечень полей данных не фиксируем жестко в коде.

Отсеивание дублей по номерам телефона

При парсинге отсеиваем объявления, которые содержат номера, которые уже ранее попадались в процессе текущего парсинга.

Загрузка Веб-страниц

Для загрузки страниц будет использован браузерный компонент на основе Playwright. Основная цель решения - осуществлять загрузку вебстраниц максимально приближенно, как это делают браузеры разных пользователей.

После загрузки должен произойти клик по плашке с телефоном и нужно дождаться загрузки телефона. Если телефон не загрузился, то выдаем ошибку и добавляем текущий элемент в конец очереди на обработку (загрузится позже). Один элемент (страница объявления или каталога) повторно обрабатываем максимум N раз (параметры конфигурации errorRetryTimes). Ключевые особенности:

- **Поддержка прокси.** Решение будет поддерживать загрузку через прокси. Прокси для минимизации блокировок будут находиться в постоянной ротации.
- **Имитация загрузки с разных устройств:** Использование системы изменения отпечатка браузера для имитации естественной активности пользователя, что значительно снижает риск блокировки со стороны веб-сайтов. Проще говоря - каждая прокся будет иметь собственный кеш браузера, куки, разрешение экрана, UserAgent и т.п. Соответственно, папка с кешем каждого браузера сохраняется отдельно.
- **Кэширование загрузки страниц объявлений (но не страниц каталога):** (важно - не путаем с кешем браузера!) Решение включает встроенное кэширование загрузки объявлений в папку, сохраняя каждую URL-страницу в отдельный локальный файл. При первом посещении веб-страницы она загружается с сервера и кэшируется. Важный момент - со страницей также кешируем телефон (который получается позже при клике на плашке с телефоном). Соответственно, учесть это при сборе данных, что телефон может загружаться как с обычной, так и с кэшированной версии. При последующих посещениях данные извлекаются из кэша. Это полезно, если вам потребуется повторно перепарсить те же страницы.

Сохранение собранных данных

Сохраняем данные в CSV файл (по ходу парсинга), который можно открыть в Excel для просмотра. Набор полей генерируем на основе текущих доступных полей сбора данных. Формат UTF8, разделитель ";"

CarModel	Phone	Name	City	Publishing	Price	URL	
Ford Edge (093) 844	Motor.we	Одеса	3 хвилини	24 900	\$	https://auto.ria.com/uk/auto_ford_edge_38819753.html	
Volvo S40 (096) 181	Олександ	Рівне	9 хвилини	7 600	\$	https://auto.ria.com/uk/auto_volvo_s40_38819749.html	
Nissan X-Trail (093) 458	Светлана	Звягель	27 хвилини	6 500	\$	https://auto.ria.com/uk/auto_nissan_x_trail_38535314.html	
Opel Vectra (067) 306	Вікторія	Гнівань	26 хвилини	2 800	\$	https://auto.ria.com/uk/auto_opel_vectra_38675687.html	
Hyundai Sonata (093) 150	Дмитро	Одеса	28 хвилини	8 700	\$	https://auto.ria.com/uk/auto_hyundai_sonata_38247519.html	
Volkswagen Golf (068) 053	Артем	По Коростиш	19 хвилини	6 650	\$	https://auto.ria.com/uk/auto_volkswagen_golf_37953098.html	
Kia Niro (067) 363	Сергій	Рівне	3 хвилини	18 700	\$	https://auto.ria.com/uk/auto_kia_niro_38656677.html	
Kia Magentis (097) 943	Eduard	Бердичів	12 хвилини	5 900	\$	https://auto.ria.com/uk/auto_kia_magentis_38819730.html	
Hyundai H-200 (096) 289	Валерій	Одеса	27 хвилини	2 500	\$	https://auto.ria.com/uk/auto_hyundai_h_200_38535258.html	
Jaguar I-Pace (067) 297	Валера	Житомир	28 хвилини	26 900	\$	https://auto.ria.com/uk/auto_jaguar_i_pace_38247301.html	
Audi A8 (068) 086	Тарас	Вікторія	27 хвилини	1 750	\$	https://auto.ria.com/uk/auto_audi_a8_38675678.html	
BMW 5 Series (073) 898	Крістіна	Немирів	20 хвилини	5 900	\$	https://auto.ria.com/uk/auto_bmw_5_series_38819696.html	
Volkswagen Tiguan (098) 806	Сергій	Ол	Рівне	35 хвилини	8 499	\$	https://auto.ria.com/uk/auto_volkswagen_tiguan_38778474.html
Volkswagen Tiguan (098) 806	Сергій	Ол	Рівне	35 хвилини	8 499	\$	https://auto.ria.com/uk/auto_volkswagen_tiguan_38778474.html
Opel Vectra (099) 393	Владисла	Рівне	20 хвилини	2 200	\$	https://auto.ria.com/uk/auto_opel_vectra_38819693.html	
Skoda Octavia (068) 500	Діма	Ладижин	28 хвилини	5 800	\$	https://auto.ria.com/uk/auto_skoda_octavia_38535223.html	
Hyundai Elantra (096) 126	Ігор	Житомир	4 хвилини	8 700	\$	https://auto.ria.com/uk/auto_hyundai_elantra_38819783.html	
Daewoo Lanos (097) 615	Людмила	Дубно	12 хвилини	2 550	\$	https://auto.ria.com/uk/auto_daewoo_lanos_38819733.html	
Mercedes-Benz C-Class (067) 895	Микола	Вінниця	28 хвилини	1 700	\$	https://auto.ria.com/uk/auto_mercedes_benz_mb_class_38247324.html	
Kia Sportage (093) 985	Александ	Одеса	27 хвилини	16 500	\$	https://auto.ria.com/uk/auto_kia_sportage_38675628.html	
Infiniti Q50 (097) 102	(Александ	Одеса	22 хвилини	22 000	\$	https://auto.ria.com/uk/auto_infiniti_q50_36730736.html	
Toyota RAV4 (098) 694	Ivan	Одеса	28 хвилини	15 000	\$	https://auto.ria.com/uk/auto_toyota_rav4_38535221.html	
Toyota Corolla (050) 298	Катя	Кропивниц	5 хвилини	8 900	\$	https://auto.ria.com/uk/auto_toyota_corolla_38819759.html	
VАЗ 2109 (067) 858	(Олександ	Острог	28 хвилини	850	\$	https://auto.ria.com/uk/auto_vaz_lada_2109_38247186.html	

Общая механика процесса

Предусмотреть возможность многопоточной работы. Также в рамках одного потока иметь возможность задать длительность паузы до запроса к следующему объявлению.

Логирование

Сохраняем в лог файл log.txt (генерируется в папке документы на компьютере)
актуальные логи

Дополнительные моменты

В идеале, чтобы в парсер было легко добавить парсинг сайта агрориа (например, за счет расширения вариантов xpath для подбора ссылок на объявлений и пагинацию, а также полей данных). Примеры ссылок:

<https://agro.ria.com/tag-kombajn/>

<https://agro.ria.com/tag-zhnivarka/>

Конфигурация

Файл конфигурации config.json (лежит в корневой директории скрипта) имеет следующий вид:

```
{
```

```
"catalogXpaths": [  
    "//section[@id='searchResults']//a[@class='m-link-ticket']",  
    "//div[contains(@class, 'ticket-item')]//a[contains(@class, 'address')]",  
    "//a[contains(@class, 'address') and contains(@href, '/auto/')]"  
,  
  "paginationXpaths": [  
    "//span[@class='page-item next']//a",  
    "//a[contains(@class, 'page-link') and contains(text(), 'Далі')]",  
    "//div[@class='pagination']//a[@class='next']",  
    "//a[@class='page-link js-next']"  
,  
  "phoneButtonXpaths": [  
    "//button[contains(@class, 'phone')]",  
    "//a[contains(@class, 'show-phone')]",  
    "//div[@class='phone_show_link']",  
    "//button[@class='show-phone']"  
,  
  "dataFields": [  
    {  
      "name": "title",  
      "xpathList": [  
        "//h1[@class='head']",  
        "//h1[contains(@class, 'heading')]",  
        "//div[@class='head-title']//h1"
```

```
        ],
    },
    {
        "name": "phone",
        "xpathList": [
            "//a[@class='phone bold']",
            "//div[contains(@class, 'seller-phones')]/a",
            "//a[contains(@class, 'show-phone')]",
            "//span[contains(@class, 'phone-number')]"
        ]
    },
    {
        "name": "name",
        "xpathList": [
            "//div[@class='seller_info_name']",
            "//div[contains(@class, 'seller-name')]",
            "//span[@class='seller-name']"
        ]
    },
    {
        "name": "city",
        "xpathList": [
            "//dd[@class='seller_info_location']",
            "//div[@class='seller_info_area']",
            ...
        ]
    }
]
```

```
    "//span[contains(@class, 'location')]",
    "//div[contains(@class, 'seller-info-location')]"
]

},
{
  "name": "date",
  "xpathList": [
    "//div[@class='footer_box']//span[contains(@class, 'date')]",
    "//span[@class='date-created']",
    "//div[contains(@class, 'publication-date')]",
    "//span[contains(@class, 'auto-date')]"
  ]
},
{
  "name": "price",
  "xpathList": [
    "//div[@class='price_value']//strong",
    "//span[@class='price']",
    "//div[contains(@class, 'price-ticket')]",
    "//div[contains(@class, 'price_value')]"
  ]
}
],
  "parsing": {
```

```
"threads": 3,  
"delayBetweenRequests": {  
    "min": 2,  
    "max": 5  
},  
"pageLoadTimeout": 30000,  
"waitForPaginationTimeout": 5000  
,  
"errorRetryTimes": 3,  
"proxy": {  
    "enabled": true,  
    "rotation": true,  
    "list": []  
},  
"cache": {  
    "enabled": true,  
    "directory": "./cache",  
    "cacheListings": true,  
    "cacheCatalog": false  
},  
"output": {  
    "file": "output.csv",  
    "encoding": "utf-8",  
    "delimiter": ";"
```

}

}