1. Определиться с функционалом

- с каких сайтов будем парсить

- какую информацию

- какой вывод информации, формат

- нужно ли делать фильтры (можно сделать поиск в том, что спарсили)

- под какую систему пишем

- внешний вид приложения

1. Кто, что будет делать?

- Дизайн

- Backend

- Frontend

Скажу честно, мне ближе Backend, но надо решать

1. Определиться с основными понятиями, объектами, переменными, что бы у Всех они были единообразны.

Мое видение (Исходя из того, что я посмотрел в видео от Дмитрия)

1. Делаем первую рабочую версию парсинга, чисто консольный вариант, как в видео с выводом в файл).
2. Дальше улучшаем, в процессе решаем и думаем над дизайном + - , и решаем где будем размещать (телеграмм бот, ПК, или сайт)
3. Дорабатываем.
4. В каждом этапе пишем тех документацию, то что делали, в принципе можно попробовать и в том же Jupyter все собрать.

Жду Ваших предложений.

**1.       Определиться с функционалом**

**- с каких сайтов будем парсить**

**- какую информацию**

**- какой вывод информации, формат**

**- нужно ли делать фильтры (можно сделать поиск в том, что спарсили)**

**- под какую систему пишем**

**- внешний вид  приложения**

**2.       Кто, что будет делать?**

**- Дизайн**

**- Backend**

**- Frontend**

**Скажу честно, мне ближе Backend, но надо решать**

**3.       Определиться с основными понятиями, объектами, переменными, что бы у Всех они были единообразны.**

**Мое видение (Исходя из того, что я посмотрел в видео от Дмитрия)**

**1.       Делаем первую рабочую версию парсинга, чисто консольный вариант, как в видео с выводом в файл).**

**2.       Дальше улучшаем, в процессе решаем и думаем над дизайном + - , и решаем где будем размещать (телеграмм бот, ПК, или сайт)**

**3.       Дорабатываем.**

**4.       В каждом этапе пишем тех документацию, то что делали, в принципе можно попробовать и в том же Jupyter все собрать.**

**Жду Ваших предложений.**

**Что такое парсинг и для чего он нужен**

**Парсинг** (**parsing**) — это сбор информации из сторонних источников и сайтов для использования полученных данных в различных целях, от аналитики до копирования, **простыми словами**, это сбор данных из различных источников. **Парсить** можно с помощью специальных программ, а также языка программирования XPath.

**Парсинг** — это процесс автоматического сбора данных и их структурирования. Специальные программы или сервисы-парсеры «обходят» сайт и собирают данные, которые соответствуют заданному условию. Простой пример: допустим, нужно собрать контакты потенциальных партнеров из определенной ниши

*Плюсы парсинга очевидны — если сравнивать его с ручным сбором и сортировкой данных:*

* *вы получаете данные очень быстро;*
* *можно задавать десятки параметров для составления выборки;*
* *в отчете не будет ошибок;*
* *парсинг можно настроить с определенной периодичностью — например, собирать данные каждый понедельник;*
* *многие парсеры не только собирают данные, но и советуют, как исправить ошибки на сайте.*

***Минусы :***

* *некоторые* [*IP-адреса*](https://kokoc.com/blog/ip-adres-chto-ehto-takoe-prostymi-slovami/) *могут быть защищены от парсинга, соответственно всю информацию собрать будет невозможно;*
* [*капча*](https://kokoc.com/terminy/chto-takoe-kapcha-pri-registracii/) *способна замедлить процесс;*
* *парсер может прервать работу, если на сайте большое количество пользователей, т. к. обновить страницу для лучшей загрузки программа не может.*

*Парсинг обычно применяют, когда нужно быстро собрать большой объем данных. Его выполняют с помощью специальных сервисов — парсеров. В этой статье мы разберем, с какой целью можно использовать парсинг, что он позволяет узнать о конкурентах и законен ли он. Также мы рассмотрим, как пошагово спарсить данные с помощью одного из инструментов.*

*Парсинг применяют для:*

1. ***Исследование рынка****. Парсинг позволяет быстро оценить, какие товары и цены у конкурентов.*
2. ***Анализ динамики изменений****. Парсинг можно проводить регулярно, чтобы оценивать, как менялись какие-то показатели. Например, росли или падали цены, изменялось количество онлайн-объявлений или сообщений на форуме.*
3. ***Устранение недочетов на собственном ресурсе.*** *Выявление ошибок в мета-тегах, битых ссылок, проблем с редиректами, дублирующихся элементов и т. д.*
4. ***Сбор ссылок, ведущих на вашу площадку.*** *Это поможет оценить работу подрядчика по линкбилдингу.*
5. ***Наполнение каталога интернет-магазина.*** *Обычно у таких сайтов огромное количество позиций и уходит много времени, чтобы составить описание для всех товаров. Чтобы упростить этот процесс, часто парсят зарубежные магазины и просто переводят информацию о товарах.*
6. ***Составление клиентской базы.*** *В этом случае парсят контактные данные, например, пользователей соцсетей, участников форумов и т. д. Но тут стоит помнить, что сбор информации, которой нет в открытом доступе, незаконен.*
7. ***Сбор отзывов и комментариев*** *на форумах, в соцсетях.*
8. ***Создание контента, который строится на выборке данных****. Например, результаты спортивных состязаний, инфографики по изменению цен, погоды и т. д.*

*Кстати, недобросовестные люди могут использовать парсеры для DDOS-атак. Если одновременно начать парсить сотни страниц сайта, то площадку можно «положить» на какое-то время. Это, разумеется, незаконно — об этом подробнее ниже От подобных атак можно защититься, если на сервере установлена защита.*

***Этапы***

*Технология парсинга работает так:*

***Этап 1.*** *Направление запроса. Пользователь заносит в программу условия соответствия выборке. Это могут быть, например, артикулы товаров, ценники, желаемый возраст лидов или их город.*

***Этап 2.*** *Парсер начинает свою работу и собирает данные с сайта. Если правильно запустить программу, то этот процесс займет буквально несколько минут.*

***Этап 3.*** *Сортирует данные по определенному признаку: буквенному, числовому и т. д. Например, программа распределяет полученные об участниках онлайн-группы данные в таблицу. В одном столбце помещает информацию о возрасте, в другом пишет город проживания, в следующем образование.*

***Этап 4.*** *Получаем отчет с данными. В отчете могут быть пометки: например, если он делал парсинг на наличие каких-либо ошибок, то они будут указаны.*

***Этап 5.*** *Выгрузка результатов. Лучше всего сохранять данные отчета в формате CSV или XML, чтобы документ можно было открыть в Microsoft Excel, Google Sheets и любом другом табличном редакторе.*

***Что могут узнать конкуренты с помощью парсинга***

*В принципе, любую информацию, которая размещена на вашем сайте. Чаще всего ищут:*

* *цены;*
* *контакты компании;*
* *описание товаров, их характеристик и в целом контент;*
* *фото и видео;*
* *информацию о скидках;*
* *отзывы.*

*Проводить такую «разведку» могут не только конкуренты. Например, журналист может провести исследование, правда ли интернет-магазины предоставляют настоящие скидки на Черную пятницу. Или искусственно завышают цены незадолго до нее и реальную цену выдают за скидку. С этой целью он может заранее спарсить цены десятка интернет-магазинов и сравнить с ценами на Черную пятницу.*

***Законно ли парсить сайты***

*Если кратко, то законно — если вы парсите информацию, которая есть в открытом доступе. Это логично, ведь так любой человек и без парсера может собрать интересующие данные. Что преследуется законом:*

* *парсинг с целью DDOS-атаки;*
* *сбор личных данных пользователей, которые находятся не на виду — например, в личном кабинете, указывались при регистрации и т. д.;*
* *парсинг для воровства контента — например, перепост чужих статей под своим именем, использование авторских фото не из бесплатных стоков;*
* *сбор информации, которая составляет государственную или коммерческую тайну.*

*Рассмотрим это подробнее с точки зрения законодательства Украины.*

*Согласно* [*ЗУ «Об информации»*](https://zakon.rada.gov.ua/laws/show/2657-12#Text)*, информация по режиму доступа делится на общедоступную и информацию с ограниченным доступом. В свою очередь информация с ограниченным доступом делится на конфиденциальную, гостайну и служебную. Определения каждого вида содержатся в* [*ЗУ «О доступе к публичной информации*](https://zakon.rada.gov.ua/laws/show/2939-17#Text)*.*

*В большей степени любой спор касательно незаконного парсинга и/или распространения информации касается именно конфиденциальных данных.*

1. *Информация о физлице, которая может его идентифицировать, априори является конфиденциальной и может быть использована только по согласию. Поэтому, чтобы парсинг был законным, парсить нужно либо деперсонифицированные данные, либо получать согласие распорядителя информации — владельца сайта, на котором зарегистрирован пользователь.*
2. *Если речь идет об информации, не являющейся персональной, она может считаться конфиденциальной, только если ее владелец определил ее как таковую. Так, чаще всего на сайтах размещается либо политика конфиденциальности, либо правила пользования сайтом. В этом документе/на этой странице указаны права и обязанности посетителей/пользователей, которые нужно соблюдать. Поэтому перед парсингом стоить проверить, не запрещен ли сбор информации и использование данных сайта.*

*Также важным является возможное нарушение авторских установленных* [*ЗУ «Об авторских и смежных правах»*](https://zakon.rada.gov.ua/laws/show/3792-12#Text) *и* [*ГКУ*](https://zakon.rada.gov.ua/laws/show/435-15)*. Перед парсингом нужно понимать, что любой тип контента защищен авторским правом с момента его создания. И только автор определяет как (платно/бесплатно), где (статья/сайт/реклама) и сколько (на протяжении срока действия лицензии/бессрочно) можно использовать его творение.*

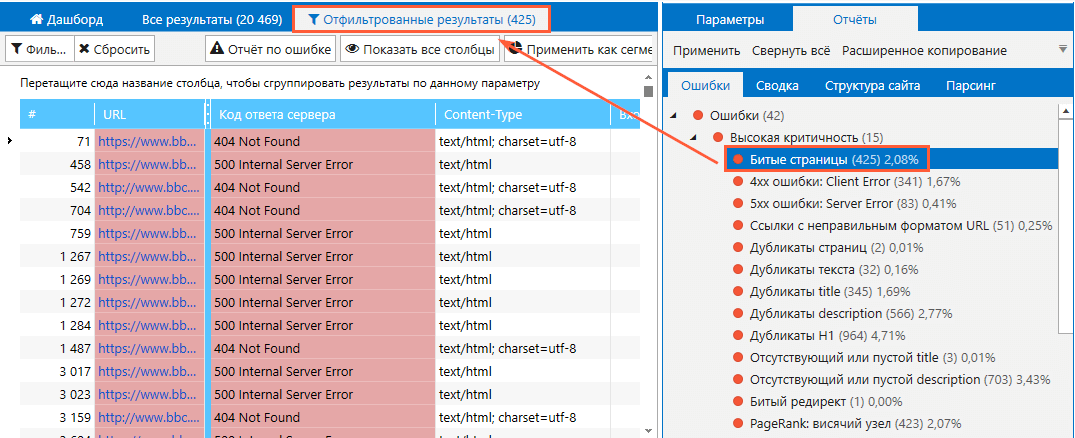
*Даже при условии правомерности парсинга, его осуществление не должно подрывать нормальную работу сайта, который парсят. Если из-за парсинга информации произойдет сбой и утечка или подделка данных, то подобные действия могут расцениваться как несанкционированное вмешательство в работу сайта, что является нарушением согласно* [*УК Украины*](https://zakon.rada.gov.ua/laws/show/2341-14/conv#n2493)*.*

*Есть еще один нюанс. Представим, что одна компания долго разрабатывала продукт, вкладывала деньги, чтобы собрать базу пользователей или покупателей, а другая спарсила все и за несколько недель создала практически аналогичный сервис или продукт. Подобные действия при наличии весомой доказательной базы могут расцениваться как нарушение условий конкуренции согласно* [*ЗУ «О защите от недобросовестной конкуренции»*](https://zakon.rada.gov.ua/laws/show/236/96-%D0%B2%D1%80#Text)*.*

***Этапы парсинга***

*Если не погружаться в технические подробности, то парсинг строится из таких этапов:*

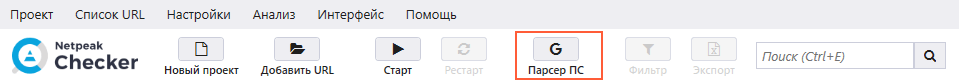
* *пользователь задает в парсере условия, которым должна соответствовать выборка — например, все цены на конкретном сайте;*
* *программа проходится по сайту или нескольким и собирает релевантную информацию;*
* *данные сортируются;*
* *пользователь получает отчет — если проводилась проверка на ошибки, то критичные выделяются контрастным цветом;*
* *отчет можно выгрузить в нужном формате — обычно парсеры поддерживают несколько.*

*Пример отчета Netpeak Spider, где критичные ошибки выделяются красным цветом.* [*Источник*](https://seo-akademiya.com/baza-znanij/servisyi/kak-polzovatsya-netpeak-spider/)

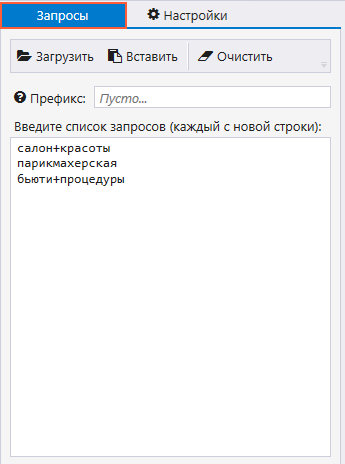
***Как парсить данные***

*Теперь давайте более подробно рассмотрим, как парсить данные. Разберем его в разрезе довольно частой задачи для менеджера — собрать базу для «холодного» обзвона. В качестве примера возьмем парсер* [*Netpeak Checker*](https://netpeaksoftware.com/ru/checker)*, с которым работаем и сами.*

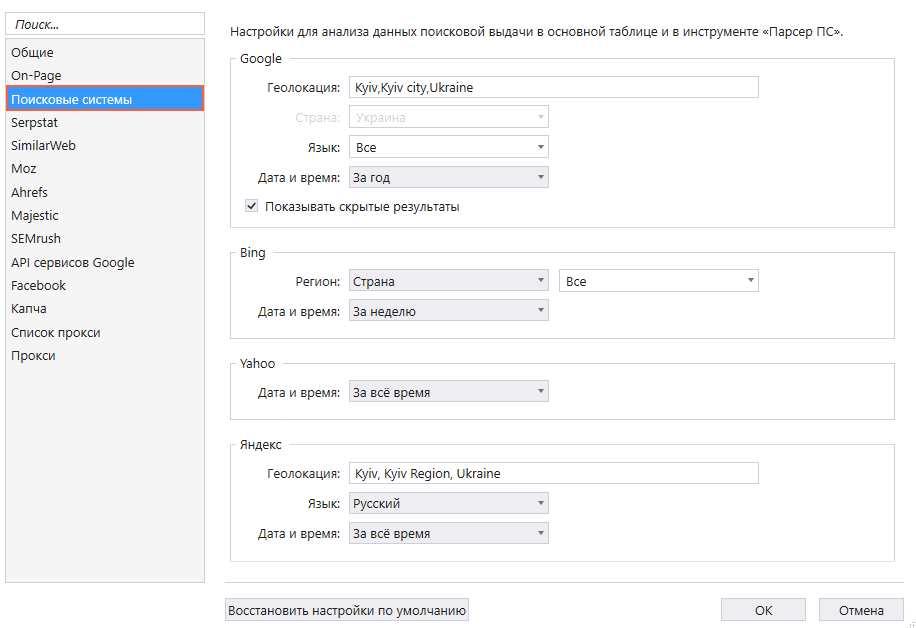
*Допустим, наша компания продает оборудование для салонов красоты. И сотруднику нужно собрать базу контактов таких компаний, чтобы позвонить и предложить им наш товар. Обычно на старте готового списка площадок у менеджера нет. Поэтому для поиска можно использовать встроенный в программу инструмент «Парсер поисковых систем».*

**

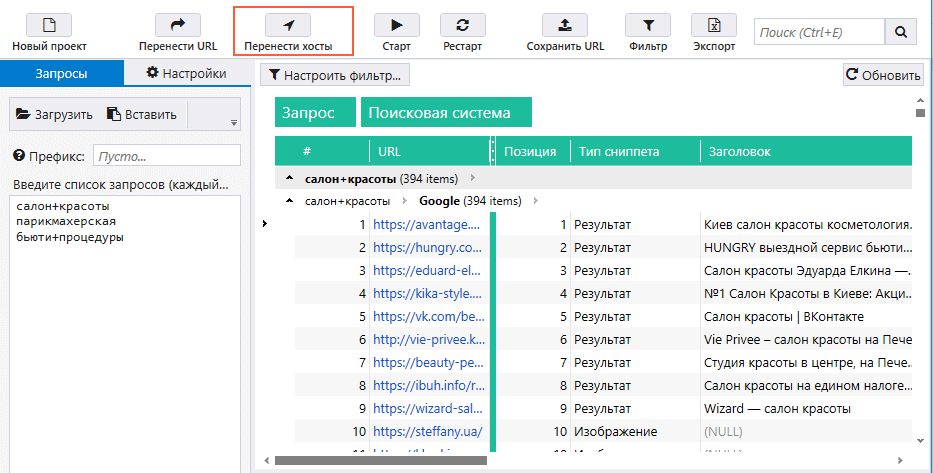
*Вводим в нем нужные запросы — «салон красоты», «парикмахерская», «бьюти-процедуры».*

**

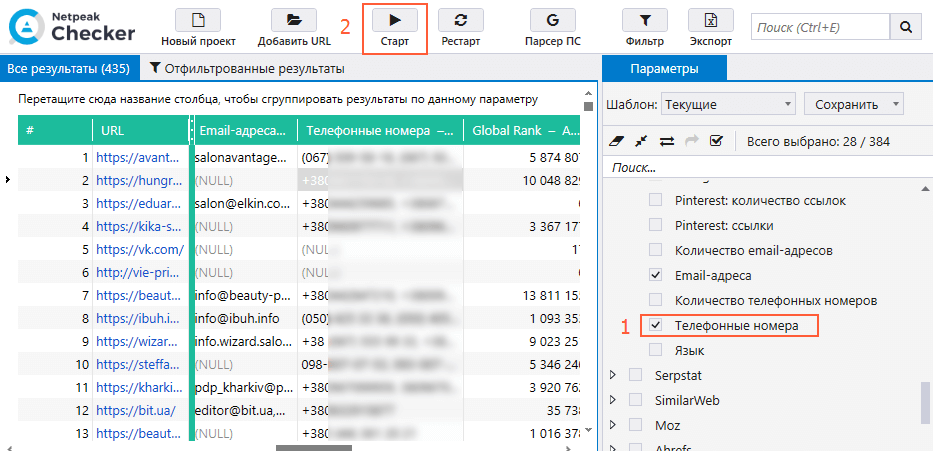
*На вкладке «Настройки» выбираем поисковую систему и количество результатов — например, топ-10 или все результаты выдачи. В дополнительных настройках указываем язык выдачи и параметры геолокации, чтобы в результаты попадали салоны красоты только из нужного нам региона. Сохраняем настройки и нажимаем «Старт», чтобы начать парсинг.*

**

*Чтобы провести парсинг номеров телефонов с главных страниц найденных сайтов, нажимаем на кнопку «Перенести хосты». После этого ссылки отобразятся в основной таблице программы.*

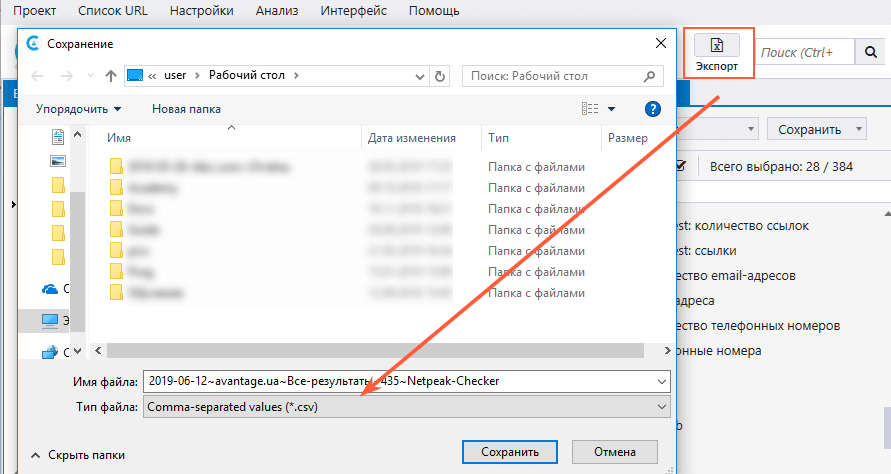
**

*Теперь, когда у нас есть полный список салонов, на боковой панели в разделе параметров «On-Page» отмечаем пункт «Телефонные номера» и нажимаем «Старт». Все найденные телефоны с сайтов и их число будут внесены в соответствующих колонках основной таблицы результатов.*

**

*Если бы у нас заранее был собран перечень необходимых адресов, мы могли бы их просто загрузить в программу и точно так же собрать телефоны.*

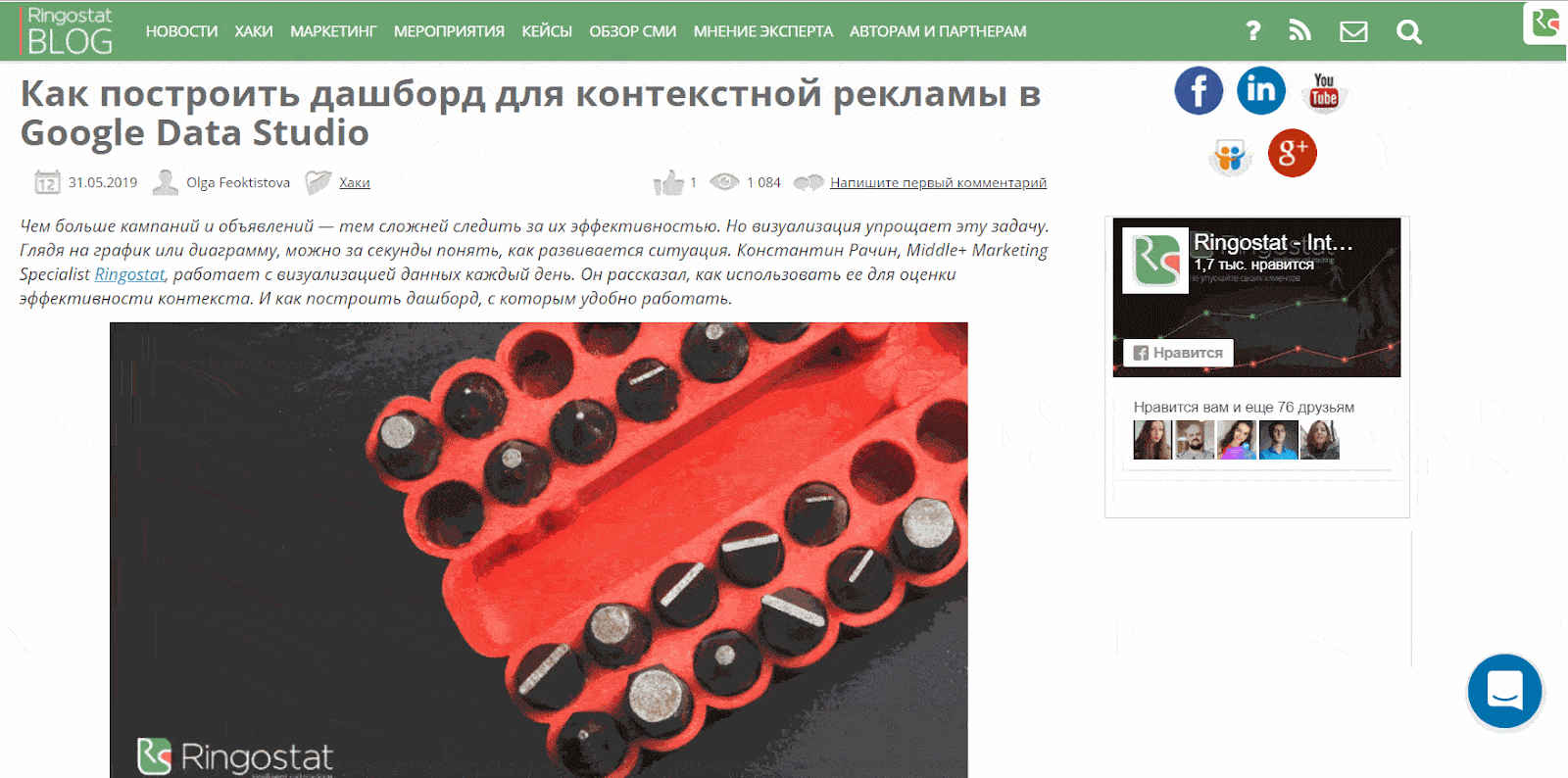
*Сохраняем данные в формате CSV, нажав кнопку «Экспорт».*

**

*Вот и все — мы получили список салонов и их телефонов.*

*Кстати, сэкономить время можно не только за счет парсинга. Вы в любом случае тратите где-то минуту, чтобы набрать номер на телефоне. Если в вашем списке хотя бы 50 компаний, на это в сумме уйдет почти час. Но есть способ тратить на набор номера одну секунду. Это* [*Ringostat Smart Phone*](https://chrome.google.com/webstore/detail/ringostat-smart-phone/pmpdmpjichjpigklkhholiejpnhcjhkf) *— умный телефон, встроенный прямо в браузер Chrome. Он позволяет звонить, просто нажав на номер, расположенный на любом сайте, в карточке CRM или просто в таблице. Как в нашем примере.*

*Подключите Ringostat, установите расширение и сможете обзвонить базу за минимальное время. При желании ее можно сразу перенести в CRM и звонить уже оттуда с помощью Ringostat Smart Phone. Тут видно, что звонок происходит мгновенно:*

**

*Этот процесс описан в статье* [*«Лайфхак для менеджера: как подготовить базу за минимальное время»*](https://blog.ringostat.com/ru/layfhak-dlya-menedzhera-kak-podgotovit-i-obzvonit-bazu-za-minimalnoe-vremya/)*.*

***Как защитить свой сайт от парсинга***

*Как мы упоминали выше, парсинг не всегда используют в нормальных целях. Если вы боитесь атаки со стороны конкурентов, площадку можно защитить. Существует несколько способов, как это сделать.*

1. *Ограничьте число действий, которые можно совершить на вашей площадке за определенное время. Например, разрешите только три запроса в течение минуты с одного IP-адреса.*
2. *Отслеживайте подозрительную активность. Если заметили сильно много запросов с одного адреса, запретите ему доступ. Или показывайте reCAPTCHA, чтобы пользователь подтвердил, что он человек, а не бот или парсер.*
3. *Создайте учетную запись, чтобы действия на сайте мог совершать зарегистрированный посетитель.*
4. *Идентифицируйте всех, кто заходит на площадку. Например, по скорости заполнения формы или месту нажатия на кнопку. Есть скрипты, которые позволят собирать информацию о местонахождении пользователя, разрешении экрана.*
5. *Скройте информацию о структуре сайта. Пусть доступ к ней будет только у администратора.*
6. *Обращайте внимание на похожие или идентичные запросы, одновременно поступающие с разных IP-адресов. Парсинг может быть распределенным. Например, через прокси-сервера.*

*В любом случае, помните, что всегда есть риск заблокировать реального пользователя, а не программу. Поэтому тут вам решать, что важнее — безопасность сайта или риск потери потенциального клиента.*

***Выводы***

1. *Парсинг — это сбор и сортировка данных с определенными параметрами. У этого инструмента масса преимуществ: скорость, отсутствие ошибок в выборке, возможность проводить парсинг регулярно. Плюс, многие парсеры не просто собирают данные, но и советуют, как исправить критические ошибки на вашем сайте.*
2. *Парсинг используется для анализа конкурентов, исследования рынка, поиска и устранения ошибок на собственной площадке, создания контента. Интернет-магазины используют его, чтобы переводить описания товаров с иностранных площадок.*
3. *Парсинг вполне законен, если вы собираете информацию, которая есть в открытом доступе. Нельзя проводить его, чтобы «положить» ресурс конкурента, украсть чужой контент или получить данные, не предназначенные для общего доступа.*
4. *Если боитесь атаки на свой сайт, парсинг можно выявить и запретить. Способов существует несколько, но многие парсеры хвастаются в сети, что умеют их обходить. Плюс, вы всегда рискуете заблокировать «живого» человека*

1. Дмитрий – backend, frontend, test
2. Анна – анализ
3. Николай – анализ
4. Евгений – frontend, test
5. Алексей – backend, test
6. анализ полученных данных по времени
7. парсить требования?
8. Создаем “0” версия продукта ( в консоле)
9. Делаем вывод в \*.CSV файл
10. Делаем фронт и переносим на сайт
11. Внутри сделать аналитику.
12. Делаем презентацию
13. Должность
14. з/п
15. город
16. скилы
17. анализ по времени (все свести в базу)???