**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное автономное образовательное учреждение**

**высшего образования**

**«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»**

**ИНСТИТУТ ИНФОРМАЦИОННЫХ**

**ТЕХНОЛОГИЙ И ТЕЛЕКОММУНИКАЦИЙ КАФЕДРА ПРИКЛАДНОЙ ИНФОРМАТИКИ**

**РЕФЕРАТ**

***Тема:***

«*Создание макросов для автоматизации веб – скрапинга с использованием VBA****»***

|  |  |
| --- | --- |
| **Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **Выполнил:**  **Стороженко Артем Владимировия**  **Студент 3 курса, группы ПИН-б-о-21-1**  **направления подготовки /специальности**  **09.03.03 Прикладная информатика**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  ***(Подпись)***  **Руководитель:**  **Елена Анатольевна Хохлова**  **кандидат философских наук, доцент, доцент кафедры прикладной информатики**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  ***(Подпись)*** |

**Ставрополь, 2023 г.**

Содержание

[**Введение** 3](#_Toc152042353)

[**Глава 1. История развития и методы веб – скрапинга** 4](#_Toc152042354)

[**Глава 2. Для чего используется web – scraping** 6](#_Toc152042355)

[**Глава 3. Создание макросов для автоматизации веб – скрапинга** 9](#_Toc152042356)

[**Глава 4. Правовой статус веб – скрапинга** 17](#_Toc152042357)

[**Заключение** 19](#_Toc152042358)

[**Список литературы** 20](#_Toc152042359)

# **Введение**

На сегодняшний день данные и информация стали одним из самых ценных товаров для бизнеса. Они представляют собой основу для проведения маркетинговых исследований, оценки конкурентов, определения тенденций и анализа цен.

Однако, автоматизированный сбор данных с веб – сайтов часто сопряжен с рядом трудностей, потому что веб – ресуры в большинстве своем предназначены для конечных пользователей, а не для удобства автоматического использования.

В связи с этим появилась необходимость в технологиях, способных извлекать со страницы информацию и сохранять для последующей обработки в таблице или файле. Такие технологии называются веб – скрейпингом.

В данном реферате будет расмотренна реализация технологии веб – скрейпинга при помощи макросов на языке программирования VBA, в частности, для извелечения табличных данных с веб – страниц на лист Excel.

# **Глава 1. История развития и методы веб – скрапинга**

История веб – скрапинга началась с создания первого веб – робота World Wide Web Wanderer, разработанного в июне 1993 году Мэтью Греем на языке программирования Perl, единственная задача которого заключалась в измерении размеров всемирной паутины. Создание этого бота стало первой реализацией концепта веб – кроулинга, то есть программы, перебирающей страницы во всемирной паутине и присваивающей ей определенный индекс, с помощью которого эта веб – страница может быть обнаружена.

В том же 1993 году в декабре была запущена первая поисковая система WWW, созданная на основе работы веб – робота JumpStation.

В 1994 была запущена первая полнотекстовая поисковая система – WebCrawler. До нее, поиск по веб - сайтам осуществлялся исключительно по их названию, теперь же, поиск затрагивал также и содержимое веб – страницы.

В 2000 году появились первые Web API и API Crawler. Их создание сильно упростило веб – скрапинг.

В 2004 году была запущена библиотека BeautifulSoup для Python. Ее главным предназначением было использование веб – скрейпинга для страниц, которые не предлагали Web API. BeatifulSoup синтаксически анализирует содержимое страниц, что помогает выделять и получать их содержимое.

Методы веб – скрапинга

1. Ручное копирование. В некоторых случаях, а именно когда веб – сайты устанавливают блокировку от веб – скрейпинга, например показ капчи или занесения IP – адреса.
2. Обращение к прокси – серверу. Если сайт представляет собой html- или xml- документ и к нему разрешены кросс – доменные запросы, то можно получить содержимое документа с помощью запроса к одному из имеющихся в Интернете прокси – сервера.
3. Синтаксический анализ HTML. Большинство веб – сайтов состоит из большого числа страниц, генерируемых динамически из основного источника – базы данных. При этом, данные одной и той же категории кодируются в похожие страницы с помощью общего скрипта или шаблона. Один из способов веб – скрапинга заключается в обнаружении таких шаблонов, извлечении их с веб – страницы и перевода в форму.
4. Document Object Model (DOM). DOM – это программа с API для HTML- И XML- документов. Она использует веб – браузер, например, элемент управления Mozilla для активации клиенстких сценариев для получения динамического содержимого страницы.

# **Глава 2. Для чего используется web – scraping**

Веб – скрейпинг может использоваться для решения множества различных задач. К его возможным сферам применения можно отнести:

1. Отслеживание цен и наличия товаров. Благодаря веб – скрейпингу, ритейлеры могут быстро и оперативно получать последнюю информацию и ценах и наличии товара у конкурентов, а также проводить начальную аналитику, представляя данные в виде наглядных таблиц и графиков. Это позволяет корректировать цены чтобы опередить конкурентов.
2. Рыночная и конкурентная разведка. Прежде чем выйти на новый рынок, очень важно изучить все его составляющие, такие как ассортимент товаров, структуру цен, объем продаж, методы маркетинга и т. д. Веб – скрапинг позволяет значительно сократить время, которое сотрудники тратят на сбор информации, позволяя сосредоточить их усилия на непосредственном ее анализе.
3. Модернизация сайтов. При создании нового сайта может возникнуть необходимость перенести на него данные со старого. Для этого часто используется именно веб – скрейпинг.
4. Мониторинг новостей. Скрейпинг может использоваться для получения информации о последних новостях путем сбора информации с новостных блогов и сайтов. Например, сервис Brand Analytics использует веб – скрейпинг для анализа потребностей клиентов и мониторинга упоминаний организации в соцесетях.
5. Анализ эффективности контента. Блогеры и SMM – специалисты используют веб – скрейпинг для анализа эффективности создаваемого контента.
6. Поиск потенциальных клиентов. Потенциальных клиентов, каким – либо образом отреагировавшим на маркетинговую коммуникацию и предоставившим свои контактные данные организации называют лидами. Основными их источниками являются заявки, оставленные через форму обратной связи на сайте организации, входящее сообщение на электронную почту и телефонные звонки. Веб – скрэйпинг позволяет собрать всю информацию о лидах из разных источников и собрать ее в одном месте для дальнейшей обработки.
7. Защита от подделок для продуктов электронной коммерции. Подделка – это мошеническая имитация надежного бренда или продукта. Оставленные без внимания, подделки электронных продуктов организации могут нанести удар по ее репутации. Используя веб – скрейпинг, маркетинговые команды и компании могут оставаться в курсе попыток подобного мошенничества и не дать им нанести вред своей репутации.
8. Настройка машинного обучения. Веб – скрейпинг может использоваться для сбора нейросетями релевантных данных для их эффективного обучения.

Преимущества веб – скрапинга

* Высокая точность. Парсеры работают автоматически, максимально точно считывая с веб – страниц информацию, полностью ислючая негативное влияние человеческого фактора.
* Рентабельность. Веб – скрапинг повышает рентабельность бизнеса, поскольку уменьшает количество персонала, который требуется для поиска информации в сети.
* Автоматизация. Веб – скрапинг обеспечивает автоматизацию задач, которые до этого выполнялись вручную, что позволяет сотрудникам освободить время от рутинных проблем и сосредоточиться на более сложной работе, а именно на анализе полученных данных. Также, это позволяет собирать данные в гораздо большем объеме, чем было бы возможно для одного сотрудника.
* Высокое качество данных. Веб – скрапинг полезен тем, что позволяет получить реальную на текущий момент информацию из Интернета, что особенно важно для исследований рынка и исследования эффективности контента.
* Стоимость. В настоящее время практически нет нужды разрабатывать собственную систему парсинга, так как на рынке широко представлены сервисы скрейпинга для различных направлений и потребностей как компаний, так и конечных пользователей, что значительно снижает стоимость внедрения этой технологии.

Недостатки веб – скрапинга

* Потребность в обновлении. Веб – сайты постоянно меняют свою структуру, а так как работа программ для скрапинга обычно сильно от нее зависит, их часто приходится обновлять.
* Защита от скрейпинга. Скрейперы всегда создают паразитную нагрузку на целевой сайт. Именно поэтому крупные порталы часто имеют те или иные системы защиты против скрейпинга: показ капчи, занесение IP – адреса в черный список, скоринг клиентов и т.д.
* Правовой статус. В процессе парсинга можно нарушить отдельные требования законов. В отдельных странах к этому относятся по-разному. Например, в России можно нарушить авторские права, коммерческую тайну и требования закона о хранении персональных данных.

# **Глава 3. Создание макросов для автоматизации веб – скрапинга**

Простейший способ автоматизации скрапинга данных в Excel –это использование метод QueryTables.Add. Он имеет следующие параметры:

* Connection. Может иметь различные типы в зависимости от вида источника данных, например для баз данных объекты ADO (ActiveX Data Objects), объект QueryTable или средство поиска данных. Но чаще всего представляет собой строку, содержащую URL адрес необходимого интернет – ресурса.
* Destination. Объект типа данных Range, обозначающий левую верхнюю ячейку активного листа, начиная с которой на листе будут размещены полученные данные.

При выполнении этой команды создается объект QueryTable, который представляет собой таблицу листа, созданную на основе данных, возвращаемых из внешнего источника.

Однако, само по себе создание таблицы запросов не означает получнеие данных из источника. После того, как таблица запросов создана, необходимо вызвать метод QueryTable.Refresh. Именно он отвечает за подключение к источнику данных объекта QueryTable, выполнение SQL – запроса и возвращением данных в указанный диапазон.

Пример простейшего скрейпера данных курсов валют с сайта ЦБ РФ с помощью VBA в Excel представлен на рисунках 1, 2.

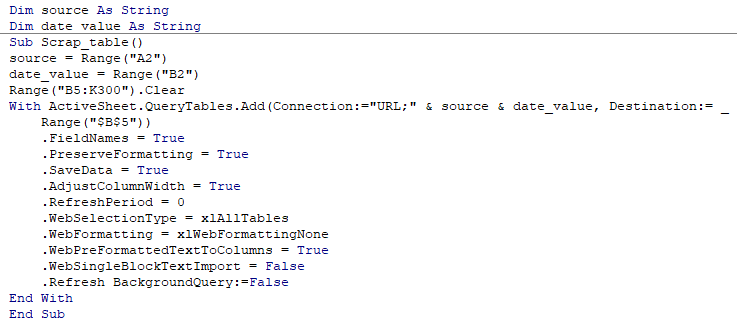


Рисунок 1 – Код скрейпера

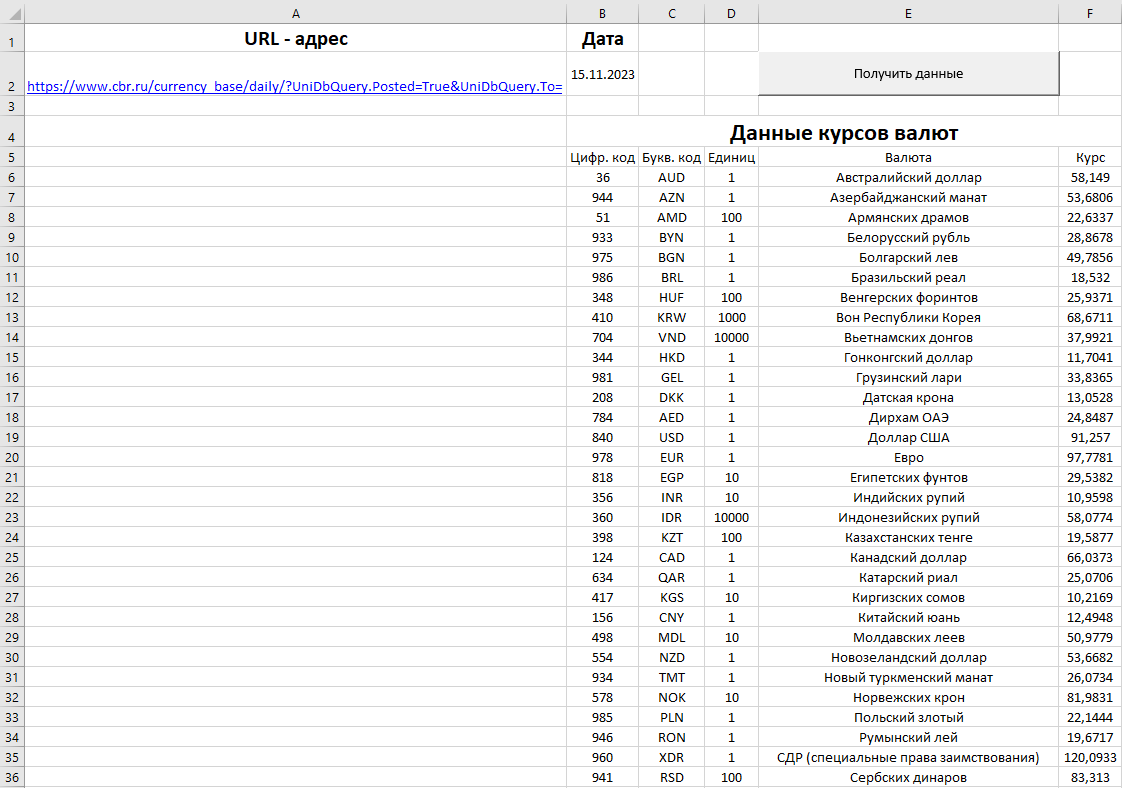


Рисунок 2 – Результат выполнения запроса

Этот скрейпер позволяет получить значения курсов валют на определенный день автоматически – достаточно лишь ввести нужную дату в ячейку B2.

Однако обычно получить данные с веб – сайта не так просто. В случае с сайтом ЦБ РФ для хранения курсов валют используется тег HTML <Table>, с чтением которого встроенные средства Excel справляются достаточно просто. Однако, на большнистве современных веб – сайтах используется другая структура таблиц, включающая в себя использование вложенных <div> тегов. Проблема такой структуры страницы заключается в том, что вложенные друг в друга теги <div> не распознаются Excel как таблицы.

Для того, чтобы решить эту проблему, необходимо воспользоваться встроенным в Excel инструментом Power Query.

Power Query – это встроенная функция в Excel, которая предоставляет возможность автоматически извлекать данные из различных источников, например с веб – сайтов, баз данных (SQL Server, Oracle, PostgreSQL и другие), а также из файлов. Также, Power Query позволяет преобразовывать данные, фильтровать их и объединять различные источники данных.

Алгоритм действий для загрузки нетабличных данных с веб – сайта с использованием Power Query выглядит следующим образом:

1. Необходимо создать запрос нужного веб – ресурса с помощью Power Query. Для этого необходимо выбрать в Excel пункты Данные → Создать запрос → Из других источников → Из Интернета. Последовательность вкладок для выбора представлена на рисунке 3.
2. Далее откроется окно, в которое необходимо внести URL – адрес необходимой страницы (рис. 4).
3. Откроется окно навигатора Power Query, в котором будут отображены все найденные на странице таблицы. Так как в этом примере мы разбираем скрапинг нетабличных данных (помещенных в теги <div>), то будет отображена лишь пустая таблица Document. Далее необходимо выбрать опцию «изменить» (рис. 5).

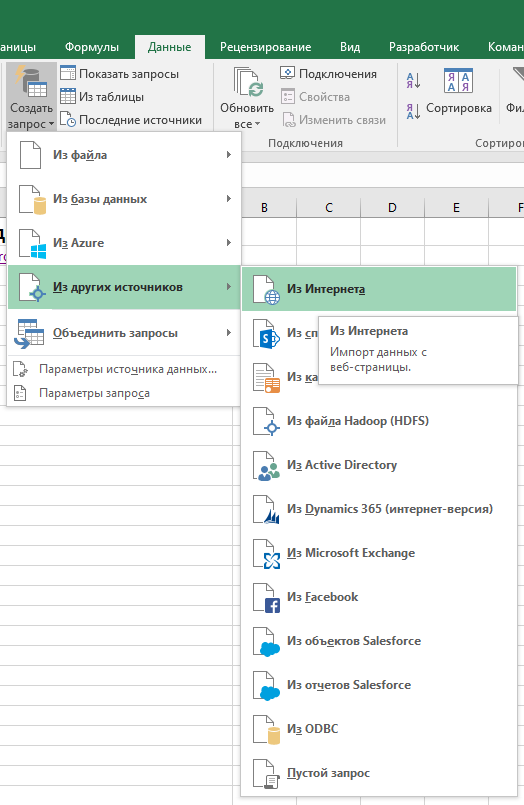


Рисунок 3 – Алгоритм запуска Power Query

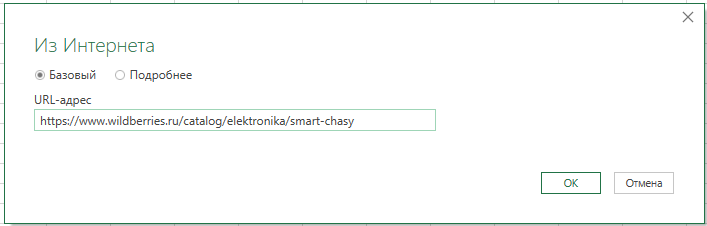


Рисунок 4 – окно для ввода URL - адреса

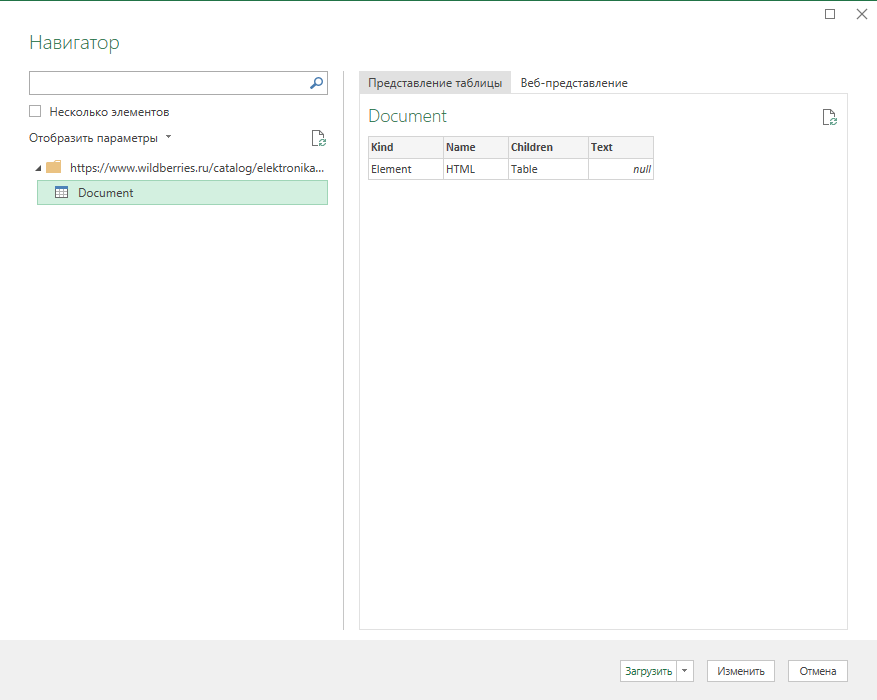


Рисунок 5 – Окно навигатора Power Query

1. Далее, необходимо получить данные веб – страницы в виде текста HTML документа. Это делается с целью поиска тегов, отвечающие за интересующие нас данные и извлечь из полученного на этом шаге текста конкретную информацию.
2. Следующим шагом нужно вырезать пробелы с левой и правой части кода HTML страницы для удобства последующей ее обработки.
3. После этого начинается непосредственное извлечение интересующих данных с веб – страницы. При использовании тегов <div> структура страницы может сильно различаться, поэтому для каждой страницы нужно подбирать свой подход. Например, на сайте книжного магазина «Лабиринт», название книг всегда описаны в свойстве “data-name”, цены – в свойстве “data-price”, а издательство – в свойстве “data-pubhouse”. Для выделения нужных строк нужно использовать текстовые фильтры (рис. 6). В результате в запросе останутся только строки с необходимой информацией (рис. 7).

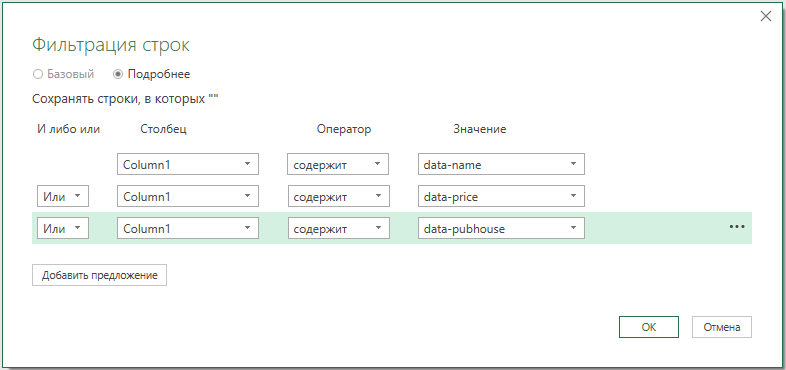


Рисунок 6 – Фильтрация строк в Power Query

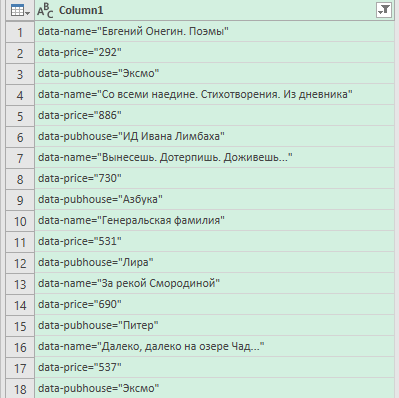


Рисунок 7 – Результат фильтрации HTML – кода страницы

1. Далее, необходимо снова обработать строки, убрав из них названия тегов, кавычки и знак “равно” (рис. 8).

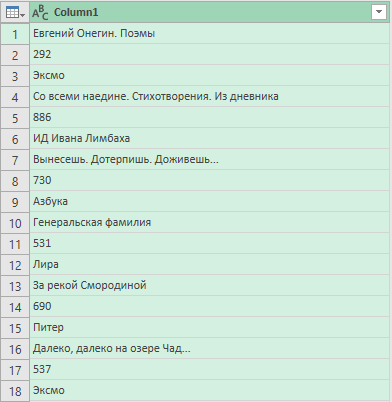


Рисунок 8 – Обработанные строки HTML документа

1. Последним шагом, нужно объединить каждые 3 строки документа в списки с помощью метода List.Split() (рис. 9).

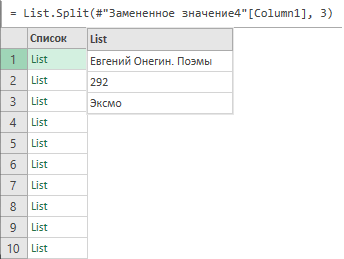


Рисунок 9 – Формула, таблица из списко, пример одного из элементов списка

В результате получится готовый запрос, который можно будет обновить с помощью макроса (рис. 10).

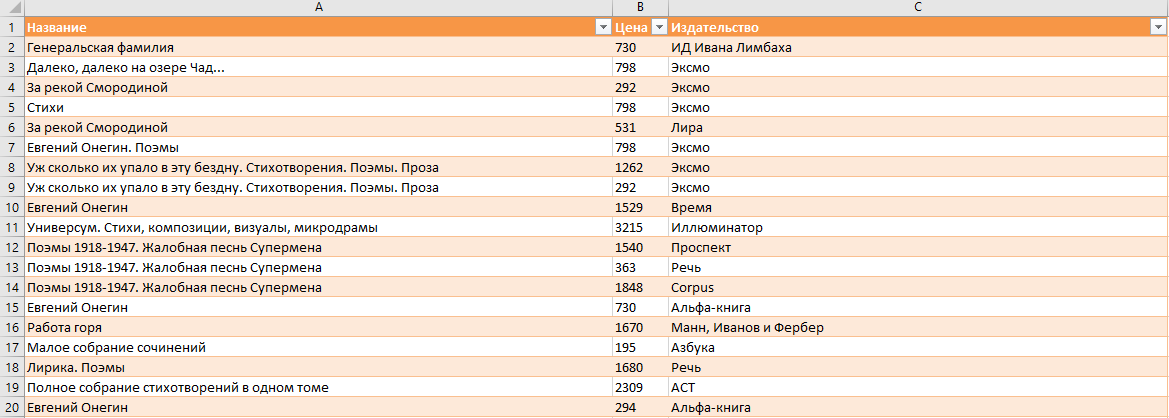


Рисунок 10 – Результат запроса

# **Глава 4. Правовой статус веб – скрапинга**

Одной из самых частых сфер применения веб – скрапинга являются социальные сети. Он постоянно используется для поиска потенциальных клиентов или новых сотрудников на таких сайтах как Facebook и Linkedin, что включает считывание информации об их контактах и персональных данных. В связи с этим у пользователей веб – скрапинга закономерно возникает вопрос, является ли такой способ его применения законным.

На первый взгляд может показаться, что если человек предоставляет свои личные данные, такие как имя, номер паспорта, адрес электронной почты, информация о работе, медицинская карта и т.д. в открытые источники, например, социальные сети, то эту информацию можно собирать любым способом и использовать в любых целях, так как она находится в открытом доступе. Однако это не совсем так. При сборе личной информации должна гарантироваться законность последующего использования такхи данных, а также соблюдаться определенные принципы при работе с информацией.

В России, скрейпинг данных из сети контролируется Федеральным законом «О персональных данных». До 1 марта 2021 года действовал пункт 10 статьи 6 этого закона, который разрешал «доступ неограниченного круга лиц к которым предоставлен субъектом персональных данных». После того, как этот пункт утратил силу, для распространения общедоступных персональных данных необходимо получение разрешения от разместившего их. Помимо этого, в пояснительной записке к этому федеральному закону указано, что изменения нацелены на запрет «третьим лицам осуществлять сбор и последующее неконтролируемое использование указанных персональных данных на интернет – сайтах в целях, отличных от первоначального их распространения». Это означает, что веб – скрейпинг социальных сетей, затрагивающий персональные данные может быть незаконным. Однако, формулировка понятия «персональные данные» сама по себе является достаточно расплывчатой: они трактуются как «любая информация, относящаяся к прямо или косвенно определенному или определяемому физическому лицу». Однако такой подход ставит под сомнение законность работы поисковых систем, работающих по схожему со скрейпингом принципу. И все же, при скрейпинге персональных данных в социальных сетях следует проявлять осторожность.

Еще одной правовой проблемой при использовании веб – скрапинга могут быть авторские права. Программа – скрейпер извлекает определенные данные с сайта и копирует их на устройство пользователя, однако как данные, размещенные на сайте, так и сам веб – сайт могут быть объектами авторского и смежного с ними прав.

В случае с авторскими правами на веб – сайт проблем не возникает, так как при извлечении с него данных путем веб – скрапинга, эти данные организуются иным способом чем на веб – сайте. При этом авторские права на информацию, размещенную на веб – сайте не являются нарушенными, если соблюдается два условия: сохранение информации носит временный характер и извлекаемая информация является неохраняемой.

В заключение, можно выделить следующие критерии законности скрапинга:

1. Нет нарушения авторских прав на материалы сайта и смежных прав на базу данных.
2. Нет обхода технических мер защиты.
3. Отсутствует недобросовестная конкуренция.
4. Отсутствуют убытки от сбоя в работе сайта.
5. Персональные данные используются в соответствии с целями их предоставления и не влекут негативных последствий для их субъекта.

# **Заключение**

Веб – скрапинг является удобной и мощной технологией для сбора данных из Интернета, а Microsoft Excel обладает самыми доступными средствами для его использования, что определяет актуальность темы этого реферата. Однако, при использовании веб – скрапинга следует проявлять осторожность при манипуляцией персональными данными и информацией, которая может быть защищена авторским правом.

# **Список литературы**

1. Eduadro Sanchez. Excel and The World Wide Web. Straight to the Point! [Электронный ресурс] URL: <https://baguzin.ru/wp/eduardo-sanches-skraping-veb-sajtov-s-pomoshhyu-excel/>
2. Что такое Power Query? [Электронный ресурс] URL: <https://learn.microsoft.com/ru-ru/power-query/power-query-what-is-power-query>
3. Парсинг нетабличных данных с сайтов [Электронный ресурс] URL: <https://www.planetaexcel.ru/techniques/13/16185/>
4. В чем разница между Web Crawling и Web Scraping? [Электронный ресурс] URL: <https://blog.froxy.com/ru/advantages-and-differences-of-web-crawling-and-web-scraping/>
5. Скрапинг интернет-ресурсов: критерии законности [Электронный ресурс] URL: <http://ipcmagazine.ru/legal-issues/scraping-of-internet-resources-criteria-of-legality>
6. Преимущества и недостатки веб – скрапинга [Электронный ресурс] URL: <https://vc.ru/u/691291-scraper-market/202951-preimushchestva-i-nedostatki-veb-skreypinga>
7. Веб – скрейпинг [Электронный ресурс] URL: <https://ru.wikipedia.org/wiki/%D0%92%D0%B5%D0%B1-%D1%81%D0%BA%D1%80%D0%B5%D0%B9%D0%BF%D0%B8%D0%BD%D0%B3>