УДК 004.624:343

ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ ПАРСИНГА ДЛЯ НАУЧНЫХ ИССЛЕДОВАНИЙ

старший преподаватель Янгличева Ю.Р. КИУ, г. Казань

**АННОТАЦИЯ**

В статье разбирается использование инструментов автоматизированного сбора данных. Демонстрируется понятие легальности парсинга с точки зрения гражданского и уголовного права Российской Федерации; определяется реализация парсинга в научных исследованиях.

*Ключевые слова:* парсинг, парсер, библиотека BeautifulSoup, Python.

**ANNOTATION**

The article examines the use of automated data collection tools. The concept of legality of parsing is demonstrated from the point of view of civil and criminal law of the Russian Federation; information is identified that goes beyond the category of public information; implementation of parsing in research.

Keywords: parsing, parser, BeautifulSoup library, Python.

Актуальность темы исследования обусловлена тем, что появилась потребность в сборе и анализе большого массива данных, а для его создания возникла необходимость автоматизировать процесс получения информации с веб-ресурсов.

В период текущей цифровизации головным источником информации стал интернет с огромным информационным ресурсом информации. Сбор и анализ необходимых данных является трудоемким процессом. Для извлечения нужной информации применяют парсеры, такого роды программы обеспечивают доступ к данным, расположенных на веб-страницах, делая процесс поиска более быстрым и эффективным. Применение данного метода в научных исследованиях открывает большие возможность работы с данными.

Законодательства Российской Федерации по правомерному использованию парсинга отражается в Федеральном законе от 27 июля 2006 г. № 149 «Об информации, информационных технологиях и о защите информации» (далее – ФЗ № 149), а именно проанализировав ст.ст. 3, 5, 7 и 8 вышеназванного нормативно-правового акта (далее – НПА), можно сделать вывод о том, что физические и юридические лица могут свободно пользоваться и распоряжаться информацией, которая принадлежит к категории «общедоступная информация» и не нарушает неприкосновенность частной жизни [2].

Они могут собирать и обрабатывать ее любым способом, если эти действия не будут ограничены нормами законодательства. Именно поэтому такой метод сбора информации, как парсинг, исходя из анализа норм данного  
НПА, может казаться на первый взгляд законным [1].

Использовать парсинг в целях ограничения здоровой конкуренции, а также для достижения злоупотребления правами субъектом или организацией на внутреннем рынке посредством установления доминирующего положения запрещено законодательством РФ.

Сбор данных весьма актуален в научных исследованиях, он является одним из самых трудоемких этапов. Парсинг — автоматизированный сбор и систематизация данных. Преимуществом его являются скорость сбора данных и, как следствие, возможность собрать большой объём информации за короткое время. Недостаток — в необходимости большого объема знаний в области программирования.

На первом этапе процесс парсинга начинается с отправки запроса на веб-ресурс и получения ответа в виде HTML кода страницы. Затем необходимо извлечь нужный текст с помощью средств парсинга, таких как библиотеки BeautifulSoup или Scrapy в Python.

На втором этапе происходит анализ полученного текста и извлечение нужных данных. Это может включать в себя поиск определенных тегов и атрибутов HTML, регулярные выражения для поиска и извлечения определенных паттернов текста, использование CSS селекторов и т.д. Полученные данные можно обработать и сохранить в нужном формате, например, в базу данных или файл.

Парсинг широко используется для сбора информации из интернета, автоматизации задач, мониторинга цен, анализа данных и других целях. Однако при парсинге следует учитывать правила использования данных на веб-ресурсе и соблюдать законы о защите персональной информации [3].

Первым шагом в разработке парсера является определение целей и функционала, которые необходимо реализовать. Затем была составлена структура кода программы, анализирующей исходный код веб-страницы и извлекающей нужную информацию.

Следующим этапом идет выбор библиотеки для парсинга HTML и работы с HTTP-запросами, такие как BEAUTIFULSOUP и REQUESTS использование этих библиотек помогут упростить процесс извлечения и обработки данных.

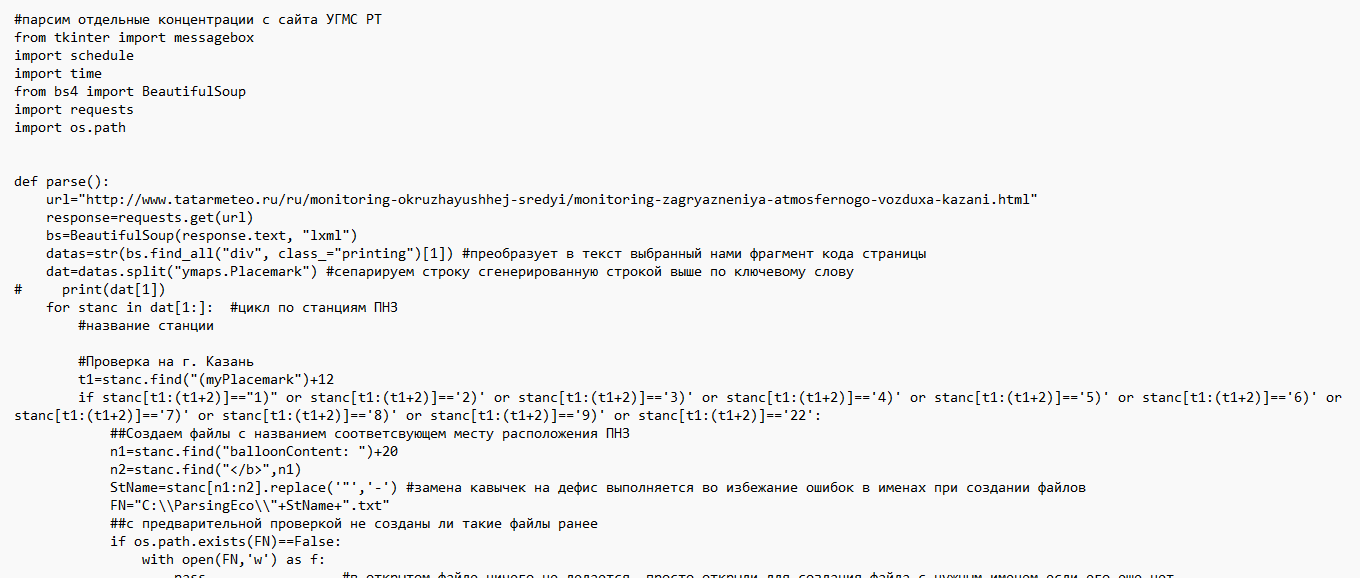
В качестве примера представлен код программы парсинга с использованием языка программирования Python. Данный парсинг позволяет собирать данные для проведения экологических исследований(рис.1).

Рис.1 Фрагмент парсинга сбора данных о концентрациях в г. Казани

Данный парсер позволит собирать данные со стационарных постов наблюдения за загрязнением атмосферы в г. Казани по схеме 4 раза в сутки сбор информации. Взяты под парсинг ряд специфических загрязняющих воздушный бассейн веществ. Станции ПНЗ привязаны по географическим координатам. В результате работы парсера будет создана база данных, которую в дальнейшем можно модифицировать в Big Data, так как данные носят структурированный характер. Собранная информация будет подвергаться статистическому анализу, выборка будет иметь крупный характер и защищает от ошибок, вызванных человеческим фактором.

Для проведения глубоких научных исследований целесообразно использовать собственные парсеры для заданной предметной области.

Литература

1. Дятлова Е.В., Янгличева Ю.Р. «Парсинг и закон»// Вестник экономики, права и социологии, 2022, № 2.
2. Федеральный закон от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и о защите информации» (с изм. и доп.) //СЗ РФ. – 2006. – № 31 (ч. I). – Ст. 3448.
3. Юлкина Н.С., Шмелев М.С., Фомина И.К. «Сравнение и анализ методов и инструментов, применяемых для сбора данных»// Молодой исследователь Дона. 2023;8(6):49–54. eISSN 2500-1779.