

ПРИМЕНЕНИЕ МЕТОДА WEB - SCRAPING ДЛЯ ПОИСКА САМОЙ ЦИТИРУЕМОЙ СТАТЬИ ПО КЛЮЧЕВОМУ СЛОВУ

Солиев И.Б.

НИ ТПУ, ОИТ ИШИТР, А1-06, e-mail: ibs2@tpu.ru

Введение

Поиск определенных тем статей, которые распространяются в научно-исследовательских журналах, был бы трудоемкой задачей, поскольку для этого потребовалось бы прочесть множество электронных библиотек или журналов, находящихся на веб-сайте [1]. Данный процесс можно эффективно выполнить с помощью применения метода web-scraping, который используется для извлечения содержимого веб-страницы в более организованные и структурированные наборы данных. В этой статье мы рассмотрим применение метода web-scraping для извлечения данных и поиска самой цитируемой статьи по ключевому слову.

Описание алгоритма

Web-scraping — это процесс автоматического сбора информации с веб-страниц. Web-scraping может быть использован для извлечения различных типов информации, таких как текст, изображения, ссылки и другие данные [2]. Web-scraping может быть полезен в случаях, когда данные не доступны в удобном формате, или когда требуется автоматизировать процесс извлечения данных.

Web-scraping основан на архитектуре клиент-сервер, при этом они взаимодействуют через интернет-соединение [3]. Веб-браузер выступает в роли клиента, где внешний интерфейс отправляет URL-адрес страницы сайта на сервер. Сервер получает запрос, а затем поддерживает пул гипертекстовых ссылок, содержащий URL-адреса выбранных веб-сайтов научных журналов. Впоследствии сервер извлекает содержимое из собранных URL-адресов, используя методы web-scraping. Затем содержимое сохраняется на сервере базы данных, которое извлекается сервером для предоставления ответов клиенту.

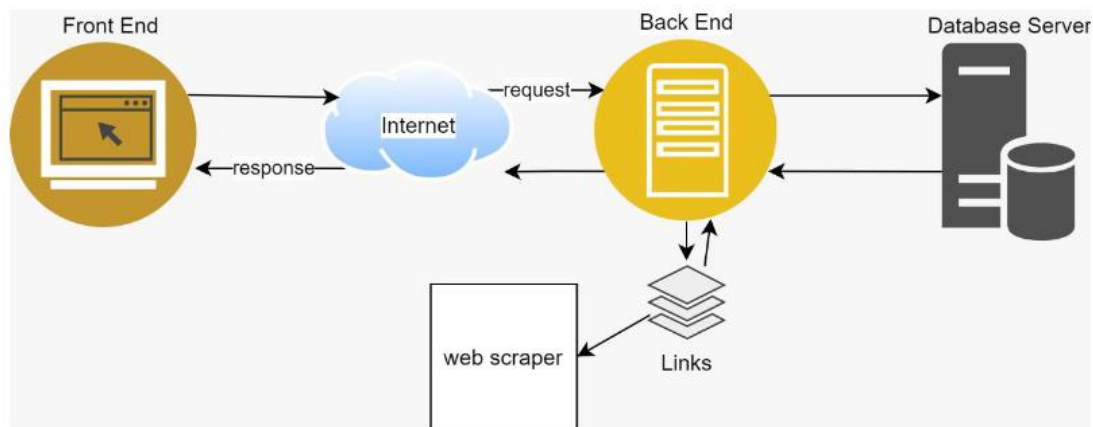


Рис. 1. Архитектура web-scraping

Применение web-scraping для извлечения данных:

Web-scraping может быть применен для извлечения данных с различных источников, включая социальные сети, новостные сайты, электронные магазины и другие веб-сайты. Для этого необходимо использовать языки программирования, такие как Python, и библиотеки, такие как BeautifulSoup и Requests [4].

Для примера рассмотрим задачу по поиску самой цитируемой статьи по ключевому слову на сайте PubMed [5]. PubMed — это база данных медицинских статей, содержащая более 30 миллионов записей.

Пример применения web-scraping для поиска самой цитируемой статьи по ключевому слову на сайте PubMed:

Для начала определим ключевое слово и количество статей, которые мы хотим рассмотреть. В данном примере мы будем искать самую цитируемую статью по ключевому слову "machine_learning" среди 100 статей.

Таблица 1

Заданные параметры входа парсинга(web-scraping) веб-сайта PubMed

Input is from:	PubMed
Please enter the keyword:	machine_learning
Please enter the number of articles:	100

Алгоритм выведет название и количество цитирований самой цитируемой статьи в PubMed по ключевому слову «machine_learning» среди 100 статей. Результат поиска представлен в таблице 2.

Таблица 2

Результат самой цитируемой статьи по ключевому слову и число цитирований

Most-cited article:	A guide to machine learning for biologists
Citation count:	2022

Заключение

Web-scraping является мощным инструментом для извлечения данных с веб-страниц, включая базы данных медицинских статей, такие как PubMed. Используя web-scraping для поиска наиболее цитируемой статьи по ключевому слову, исследователи могут быстро и легко определить наиболее важные и актуальные работы в своей области исследования.

Список использованных источников

1. Yesi Novaria Kunang, Susan Dian Purnamasari, et al. 2018. Web scraping tech-niques to collect weather data in South Sumatera. In 2018 International Conference on Electrical Engineering and Computer Science (ICECOS). IEEE, 385–390.
2. Soliev Iskandar // Approaches to knowledge extraction from scientific text // Студенческий вестник (Общество с ограниченной ответственностью "Интернаука") Номер: 47 (196), Москва: 2022 - С. 65-69.
3. D Pratiba, MS Abhay, Akhil Dua, Giridhar K Shanbhag, Neel Bhandari, and UTKARSH SINGH. 2018. Web Scraping And Data Acquisition Using Google Scholar. In 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS). IEEE, 277–281.
4. Ryan Mitchell. 2018. Web scraping with Python: Collecting more data from the modern web. " O'Reilly Media, Inc."
5. Web-scraping-method-to-extract-journal-information-from-PubMed [Электронный ресурс]. – URL: <https://github.com/IskandarAs/Web-scraping-method-to-extract-journal-information-from-PubMed.git>