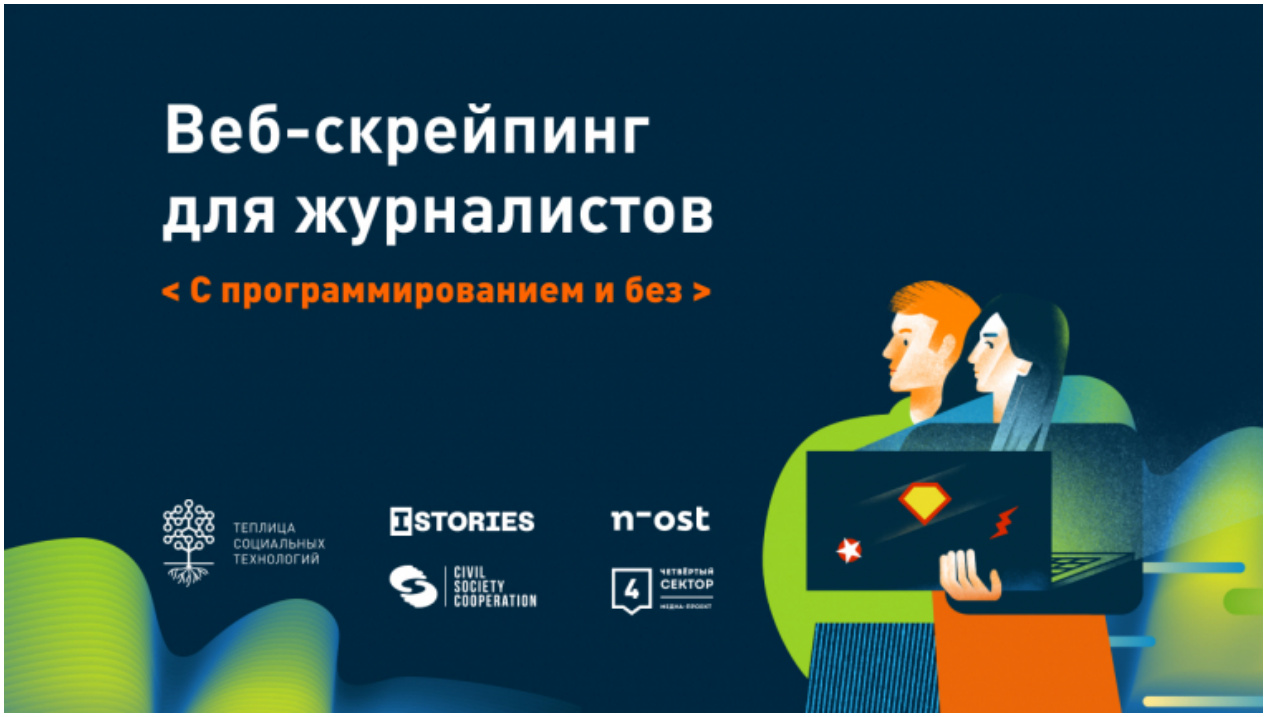
Теплица социальных технологий

Материалы курса «Веб-скрейпинг для журналистов. С программированием и без»

Алиса Цветкова

https://te-st.org/reports/web-scraping-course/

Статья обновлена 31 августа 2023



Команда Теплицы публикует конспект с онлайн-курса о веб-скрейпинге для журналистов. Эксперты рассказывают о том, как можно применять веб-скрейпинг для дата-исследований с использованием программирования и без него. Онлайн-курс провела команда фонда n-ost вместе с Теплицей социальных технологий, «Четвертым сектором» и «Важными историями».

Ведущие курса: редакторка дата-отдела издания «Важные истории» Алеся Мароховская, дата-журналист и эксперт по визуализации данных Андрей Дорожный и журналист медиапроекта «Четвертый сектор» Михаил Данилович.

Вебинар № 1: Веб-скрейпинг для журналиста

Первый вводный вебинар помогает понять, в каких случаях нужно ли использовать программирование для веб-скрепинга, а когда можно обойтись другими инструментами (о них Андрей Дорожный подробнее рассказывает на втором вебинаре).

Веб-скрейпинг — это технология получения веб-данных путем извлечения их со страниц веб-ресурсов. Это такой инструмент, который позволяет «поскрести по сусекам», когда сложно дотянуться до какой-то проблемы или до необходимых данных обычным способом.

Прежде чем извлекать данные с помощью веб-скрейпинга, убедитесь, что:

- необходимых вам данных нет в открытом доступе;
- на них нельзя сделать запрос (например, государственные данные);
- на сайте, с которого вы бы хотели их извлечь, нет API.

Так как веб-скрейпинг довольно ресурсозатратная процедура, убедитесь, что она действительно нужна. Далеко не во всех случаях можно сделать скрейпинг сайта с помощью автоматизированных инструментов. Использовать программирование стоит, если:

- сайт сделан, как картинка, представлен в формате PDF;
- сайт неструктурированный;
- на сайте есть динамические данные;
- на сайте есть paywall (ограниченный бесплатный доступ к сайту);
- на сайте есть капча.

Презентация вебинара.

Вебинар № 2: Скрейпинг данных без программирования

Практический вебинар Андрея Дорожного. Эксперт рассказывает о сервисах, которые помогут журналисту собирать данные с сайтов автоматически и без программирования. Для занятия понадобится браузер Google Chrome или Яндекс.Браузер. Необходимо также скачать несколько расширений для браузера: первое, второе, третье, четвертое.

Вебинар № 3: Скрейпинг с Python (часть 1)

Алеся Мароховская рассказывает, как использовать библиотеки BeautifulSoup и Requests, чтобы скрейпить сайты с помощью Python.

Для участия в вебинарах Алеси нужно иметь хотя бы базовые знания языка программирования Python. Также нужно установить Anaconda, чтобы на занятии вы могли использовать Jupyter Notebook (идет в комплекте Anaconda). Скачать дистрибутив можно по ссылке.

Можно использовать любой другой редактор кода, если он кажется более удобным, но на занятии Алеся использует Jupyter Notebook.

Вебинар № 4: Скрейпинг с Python (часть 2)

Алеся Мароховская рассказывает, как использовать библиотеку Selenium для более сложных случаев скрейпинга с помощью Python.

Для участия вам понадобится все тот же Jupyter Notebook (или другой редактор кода). Вам также нужно будет обновиться до самой последней версии браузера Chrome. И скачать такую же версию Chrome Driver по ссылке. Обязательно убедитесь, что у браузера и Driver одинаковые версии.