UK Data Service

Web-scraping for Social Science Research: Websites as a Source of Data

Dr Diarmuid McDonnell
Research Associate

23 April 2020

New Forms of Data Training Series

Upcoming webinars:

- Web-scraping for Social Science Research: APIs as a Source of Data (30 April 2020)
- Being a Computational Social Scientist (12 May 2020)

Upcoming coding demonstrations:

- Introduction to Python for social scientists (06 May 2020)
- Collecting data I: web-scraping (13 May 2020)

Past webinars:

- Getting Data from the Internet (16 April 2020)
- Web-scraping for Social Science Research: A Case Study (27 March 2020)

UK Data Service

Table of Contents

What is web-scraping?

It is a computational technique for capturing information stored on a web page.

It is generally implemented using a programming script, although there are software applications that you can use.

It is relatively simple to implement using open-source programming languages e.g., Python, R.

Why collect data from the web?

Web pages can be an important source of publicly available information on social phenomena of interest.

Web pages can store a range of different data types including files, text, photos, videos, lists etc, all of which may be collected and marshalled for research purposes.

Once collected, data can be reshaped into a familiar format (tabular) and linked to other sources of social science data.

Logic of web-scraping

We need to know the following:

1. The location (i.e., web address or URL) where the web page can be accessed. For example, the UK Data Service homepage can be accessed via https://ukdataservice.ac.uk.
2. The location of the information we are interested in within the structure of the web page. This involves visually inspecting a web page's underlying code using a web browser.

Then we need to do the following:

3. Request the web page using its web address.
4. Parse the structure of the web page so your programming language can work with its contents.
5. Extract the information we are interested in.
6. Write this information to a file for future use.

UK Data Service

What is the value of web-scraping for social science research?

Web-scraping is a mature computational method, with lots of established packages (e.g., `requests` and `BeautifulSoup` in Python), examples and help available.

Using computational, rather than manual, methods provides the ability to schedule or automate your data collection activities.

The richness of some of the information and data stored on web pages is a point worth repeating.

Computational methods not only enable accurate, real-time and reliable data collection, they also permit the reshaping of data into familiar formats (e.g., a CSV file, a database, a text document).

UK Data Service

What are the limitations of web-scraping for social science research?

Web-scraping may contravene the Terms of Service (ToS) of a website.

Lack of certainty around the legal basis of web-scraping.

Web pages are frequently updated, therefore changes to their structure can break your script. It can be a lot of work maintaining your code, especially if you make it available for use by others.

Some websites may be advanced enough that they throttle or block scraping of their contents.

Web-scraping, and computational social science in general, is dependent on your computing setup.

What are the ethical implications of web-scraping for social science research?

First-and-foremost, web-scraping is a component of your research project, which itself must receive ethical approval from your institution.

The impact of web-scraping on the data owner's website:

- Each request you make to a website consumes computational resources, on your end and theirs;
- Web-scraping, especially frequently scheduled scripts, can overload a server by making too many requests, causing the website to crash;
- Individuals and organisations may rely on a website for vital and timely information, and causing a website to crash could carry significant real-world implications.

Questions

Dr Diarmuid McDonnell

diarmuid.mcdonnell@manchester.ac.uk

Further resources and help

Repository: https://github.com/UKDataServiceOpen/new-forms-of-data

Youtube: https://www.youtube.com/user/UKDATASERVICE

Help: ukdataservice.ac.uk/help/

Subscribe to UK Data Service news at https://www.jiscmail.ac.uk

@UKDataService

UKDataService

UK Data Service