UWS UNIVERSITY OF THE WEST of SCOTLAND

# Web-scraping for Social Scientists

2024-06-05

Dr Diarmuid McDonnell

SGSSS Summer School 2024

# Outline

1. The value, logic and practice of web scraping
2. Understanding the structure of web pages
3. Extensions
4. Conclusion

UWS UNIVERSITY OF THE WEST of SCOTLAND

## 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.

UWS UNIVERSITY OF THE WEST of SCOTLAND

Computational is the key difference: copy-and-pasting information from web pages is usually allowed (or undetectable!), though the manual approach carries considerable disadvantages in terms of accuracy and labour resource.

I always advocate the flexibility of writing your own code (nevermind the intellectual benefits that accrue from learning such skills), however the ultimate aim is to collect data, so there are out-of-the-box solutions you can avail of e.g., Excel, Scrapy.

You do not need to be highly computationally literate, nor write screeds of code: this is a popular and mature computational method, with tons of documentation

and examples for you to learn from.

## 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 structure (tabular) and linked to other sources of social science data.

UWS UNIVERSITY OF THE WEST of SCOTLAND

Coming from a social research perspective here, though it is of course commercially valuable also e.g., Google or price comparison sites.

However, the data stored on websites are typically not structured or formatted for ease of use by researchers: for example, it may not be possible to perform a bulk download of all the files you need (think of needing the annual accounts of all registered companies in London for your research...), or the information may not even be held in a file and instead spread across paragraphs and tables throughout a web page (or worse, web pages). Luckily, web-scraping provides a means of quickly and accurately capturing and formatting data stored on web pages.

# What is the 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 BBC homepage can be accessed via https://bbc.co.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.

UWS UNIVERSITY OF THE WEST of SCOTLAND

# What is the logic of web scraping?

Then we need to **d**o 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.

UWS UNIVERSITY OF THE WEST of SCOTLAND

# What is the value of web-scraping?

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.

Collect data at scale (more concerned with coverage than sampling).

Web scraping can be an accurate and reliable data collection method.

UWS UNIVERSITY OF THE WEST of SCOTLAND

# What are the limitations/challenges?

"Data on the web typically does not come in a format amenable to analysis." (Hogan, 2022: 78)

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 is dependent on your computing setup.

Some ethical and legal complications that must be navigated/avoided.

UWS UNIVERSITY OF THE WEST of SCOTLAND

# What is a web page?

It is a document which can be displayed in a web browser (e.g., Firefox, Safari etc).

A **website** is a collection of web pages that are connected in various ways.

A **web server** is a computer that hosts/stores a website on the Internet.

A **URL (uniform resource locator)** is the location of a web page on the internet. (Mozilla, 2021)

UWS UNIVERSITY OF THE WEST of SCOTLAND

## How are web pages structured?

Web pages are written in a language called **Hyper Text Markup Language (HTML)**.

HTML describes the nested structure of a web page.

HTML consists of a series of elements, which are distinguished using tags.

HTML elements tell the browser how to display the content (e.g., fonts, colours, sections).

UWS UNIVERSITY OF THE WEST of SCOTLAND

Markup languages typically use tags to open and close levels of the hierarchy.

Tags enclose data ("values") and also have attributes (e.g., fonts, emphasis, ids). There are also self-closing tags that can have useful attribute information.

Best way to learn is to examine a web page's structure.

# Exercise

**Mary's Meals**

https://www.marysmeals.org/what-we-do/our-impact

Using the six steps from earlier, write a solution for scraping information about how many children are fed every day by this charity.

UWS UNIVERSITY OF THE WEST of SCOTLAND

Remember we need to know two pieces of information, and do four steps.

# Extensions and Considerations

*Application Programming Interfaces (APIs)* = online databases that are designed for making requests to.

*Maps* = Maps embedded with data that could be scraped (e.g., organisations in your area).

UWS UNIVERSITY OF THE WEST of SCOTLAND

# Questions and Comments

UWS UNIVERSITY OF THE WEST of SCOTLAND