UWS UNIVERSITY OF THE WEST *of* SCOTLAND

Fundamentals of Web-scraping

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 scrape 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.

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.

What is the logic of web-scraping?

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.

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.

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

What are the downsides 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.

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

Questions and Comments

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.

(Mozilla, 2021)

How are web pages structured?

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

HTML describes the 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.

Exercise

Mary’s Meals

https://www.marysmeals.org/who-we-are/how-we-spend-donations

Using the six steps from earlier, write a solution for scraping information about how donations are spent, and the aim’s of the charity.

Questions and Comments