introduction

August 28, 2024

1 Webscrapping

For Webscrapping basic knowledge of web technologies is necessary. So we start with a short introduction.

1.1 Introduction

1.1.1 Basic Web Knowledge

A simple Website (e.g https://hsg.crbn.ch) consists of HTML pages. These are normal text files (like `.py` files) that contain some (text-)content together with markup defining how to display the content (e.g. as Headers, Paragraphs or Lists).

A HTML page stored locally can be opened directly in the browser. If the page uses links, images, or style-sheets, these are just pointers to other files. If these files exist at the right location locally, the page will render correctly without internet access.

To make such a website available for others, one needs to put the files on a webserver. A Webserver is - a computer reachable from other computers on the interent - with a programm handling specific messages sent to the computer (requests)

When a user request a website from a server, the server just sends the requested file(s). The users browser then stores the file locally (in memory or on disk) and we are back at situation above.

Modern browsers come with Developer Tools that allow you to look at the things happending behind the scenes. Especially the Inspector and Network tabs are useful when scrapping data from websites.

1.1.2 wget

Since any website must be stored locally before it can be displayed, there are programs that allow you to download a full website for offline viewing/processing.

`wget` is the unix tool of choice for this. The magic invocation to download the full `hsg.crbn.ch` is

```
wget -e robots=off -r -p -k -np --restrict-file-names=ascii,unix --adjust-extension https://hsg
```

(Look at the manual to see what different options do.)

1.1.3 JavaScript

Unfortunately modern websites are not as simple as the explanation above. Instead of just providing content they also contain programms executed by the browser to load additional data and modify

the displayed page.

This language is called JavaScript and can make webscrapping significantly more complex.

One way to see JavaScript in action is to load a modern https://www.nzz.ch/, open the Network tab in the Developer Tools and then scroll down. At some point more articles will be loaded (new lines appear in the Network tab) and added to the page.

1.1.4 Resources for Background

There are lots of resources around the internet. Two possible starting points to learn web technologies are * https://developer.mozilla.org/en-US/docs/Learn/Getting_started_with_the_web * https://www.w3schools.com/html/

1.2 Examples

The example folder contains a collection of examples that demonstrate different options to extract data from websites. The recommended order is * Extracting Data from Wikipedia * Using SEC Data from Edgar * Fetching Data from Suchtindex * Accessing Google Patents * Interacting with Websites using Selenium