Advanced webscraping with Selenium

Etienne Bacher

Description

Webscraping consists in using programming tools to extract content directly from webpages. This is more and more used to build original datasets based on information that isn't stored anywhere else. Most of the time, the website that one wants to scrape is static, meaning that if you give the same URL address to several people, they will see the same thing on the webpage.

However, some websites are dynamic, meaning that the actions that one makes on the webpage (clicking on a button, selecting an input, etc.) do not change anything to the URL address. Giving the same URL to different people can then lead to different results if one person performed some actions and another didn't.

This makes webscraping harder, because we need to know how to mimic one's actions on a webpage in order to extract the content that we need.

This is where tools like `Selenium` come into action. `Selenium` is a tool that allows one to replicate their browser actions from the command line. It is therefore possible to create a program that will open a browser, click on a button, download files, etc.

The goal of this training is to familiarize PhD students, postdocs and researchers with `Selenium`.

Pre-requisites

We can use `Selenium` through different languages, such as `R` and Python. This training will be made in `R`, using the package `RSelenium`, but the methods and functions used should be easily convertible in Python.

This training will focus on the learning of `Selenium`. Therefore, some of the following skills and software are needed.

`R`

Need to know how to:

- install and load packages;
- write lists and vectors;
- write `for` loops.

Preferred but not required:

- familiarity with the core `tidyverse` packages (`dplyr`, `tidyr`);
- familiarity with the package `rvest`;
- familiarity with webscraping;
- know how to write custom functions.

Software

Make sure that you installed the package `RSelenium` before the training. This may require the installation of Java, which itself can require asking IT.