Introduction to Web Scraping

DIW Berlin

Kevin Tran
University of Bristol
kevin.tran@bristol.ac.uk

September 01, 2022, 09:30 - 11:00, 11:15 - 12:45, 14:00 - 15:30
September 02, 2022, 14:30 - 16:00

1 Description

This short course is meant to give an overview over the most common web scraping techniques. The idea is to have an interactive course in which the participants get their hands on actual code and work with it. Therefore, please bring your own computers, if possible. The main aim is to cover several approaches that are needed to scrape different types of data from different websites. In the end, participants should have an idea of how to approach the task of web scraping any website they are interested in. As an exercise spanning the entirety of the course, participants are encouraged to choose a website that they are interested in and try to build a scraper using the codes and knowledge they gain during the course.

The codes for the course are written in Python. The course also includes a very short introduction to Python but due to the limited time, we will not be able to cover all the Python concepts needed. Therefore, it would be helpful if you look at some preparatory material to get somewhat familiar with the language. Most importantly, please take the time to install a Python distribution on your computer and some of the packages that we will need for the course.

The course is split into four 90-minute sessions over two days. On day 1, we will cover the bulk of the material over three session. First, we will cover a short introduction to Python and some basic web scraping concepts. Then, we will look at how to gather data if an Application Programming Interface (API) is available. Finally, we will cover techniques for retrieving information from HTML code such as HTML parsing and text pattern matching. On day 2, we will look into browser automation, a technique that is used to scrape websites that load dynamically. Finally, we will leave some time to discuss issues with your own scraper.

The time gap between the last session of day 1 and the session on day 2 is meant to give you some time to work on your own scraping projects if you so wish.

2 Prerequisites

No prior programming experience is required to follow this course. The course includes a very short introduction to Python. Nevertheless, it might be easier for you to follow if you know some basic concepts of Python. The following tutorial covers these basic concepts: `https://www.w3schools.com/python/default.asp`

- Get familiar with the syntax of Python (Link)
- Know how a function looks like in Python (Link)
- How to use packages (Link)
- How to read and write files (Link)

3 Further reading

- Virtual environments (Link)
- Cookiecutter Data Science Project Template (Link)

4 Schedule

- September 01
  - 09:30 - 11:00
    - Very short introduction to Python
    - Basics of webscraping
  - 11:15 - 12:45
    - APIs
    - HTML parsing I
  - 14:00 - 15:15
    - HTML parsing II
    - Text pattern matching
- September 02
  - 14:30 - 16:00
    - Browser automation (Selenium)
    - Questions, Troubleshooting of own code