BCCP Web Scraping Course

Julian Harke and Kevin Tran[1]

June 24, 2019, 09:30 - 12:30
June 25, 2019, 09:30 - 12:30
June 26, 2019, 14:00 - 17:00

1 Description

This short course is meant to give an overview over the most common webscraping 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, we propose that participants can 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 familiar with the language a bit. In particular, please also 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 up in three half days. On day 1, we will cover a short introduction to Python and some basic webscraping concepts. Then, we will look at how to gather data if an Application Programming Interface (API) is available. On day 2, we will cover techniques for retrieving information from HTML code such as HTML parsing and text pattern matching. On day 3, we will look into browser automation, a technique that is necessary in particular to scrape websites that load dynamically. Finally, we will leave some time to discuss issues with your own scraper.

2 Prerequisites

No prior programming experience is required to follow this course. We will give you a very short introduction to Python. Nevertheless, it will 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)

[1] Julian is a Research Fellow at the WZB and a PhD student at Vrije Universiteit Amsterdam. Kevin is a PhD student at the DIW Graduate Center and Technische Universität Berlin. Both are members of the Berlin Centre for Consumer Policies (BCCP).

3 Further reading

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

4 Schedule

- June 24: 09:30 - 12:30
  - Very short introduction to Python
  - Basics of webscraping
  - APIs
- June 25: 09:30 - 12:30
  - HTML parsing
  - Text pattern matching
- June 26: 14:00 - 17:00
  - Browser automation (Selenium)
  - Questions, Troubleshooting of own code

If you want to join the masterclass, please register with Juliane Metzner (jmetzner@diw.de).