Hands-On Introduction to Data Scraping

With increasing availability of digital data, the buzz word ‘big data’ is commonly used to describe information about online environments like web pages and social networks. While such data is relevant for a variety of research areas, getting and processing it – data scraping - is a common hurdle for social scientists. This course offers a hands-on and systematic introduction to data scraping using the open-source programming language Python. Participants will learn automated methods for retrieving data from Application Programming Interfaces (API) such as Twitter, as well as from web pages and PDF files. After successfully completing this course, participants will be able to independently conduct data scraping projects for social science research.

No prior knowledge about programming with Python is necessary. However, participants will get the most out of this course if they already have experience with Syntax based coding with software like Stata or preferably other programming languages like R.

The course will be offered in 2 blocks á 2 x 8 hours (i.e. 4 full days). Attendance of all four days is obligatory. After the first block participants are expected to work on a take-home exercise to make sure they acquired the necessary skills for the second block.

Block 1: Introduction to Programming with Python (Day 1 / 2)

- Organizational matters
- Introduction to the programming environments (Python, Jupyter Notebooks)
- Python Basics (numeric and sequential data types, functions, loops)
- Text as data (string manipulation, regular expressions)
- Hands-on exercises

Block 2: Data Scraping (Day 3 / 4)

- Foundations of web technologies (HTML/XML, JSON)
- Web Scraping (extraction of blog articles, working with PDFs)
- API's (Twitter, maybe Facebook - thanks Cambridge Analytica!)
- Hands-on tasks
- Recap of what we learned and outlook for skills we were not able to cover

Lecturer: Carsten Schwemmer, Universities of Bamberg & Konstanz

Literature:

- Downey, A. (2015). Think Python: How to Think Like a Computer Scientist (2nd ed.). O'Reilly Media, Inc..
- Mitchell, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web (1st ed.). O'Reilly Media, Inc..
- Russell, M. (2013). Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, And more (2nd ed.). O'Reilly Media, Inc..