Web Scraping for Scientists: An Introduction with Python

This version: (First official draft)

Course instructor

Name: **Prof. Dr. Jens Foerderer**
Room: L230, Heilbronn Campus
Tel.: +49 7131 264 18 802
Mail: office.cdt@wi.tum.de
Web: https://www.wi.tum.de/prof-dr-foerderer

Application procedure

Goal and target audience
Doctoral students and post-docs in business.

Application process
Please send an email to the above-stated email address with a registration request that includes your name (see below).

Application deadline: November 1, 2022

Prerequisites

Web Scraping: This is a beginner's course in web scraping. There is no need to be familiar with any concept of web scraping. We will start from scratch.

Python programming language:
We will apply the programming language Python to develop web scrapers during the course. Course participants are required to be familiar with the very basic concepts of Python programming. There is no need to be an expert in Python. You are familiar with Python if you have a Python environment set up and are able to answer the following questions: What is a method, what is a class? What are lists, conditions, and loops? What are packages, and how do I use them?
If you are not yet familiar with Python, then I recommend taking the following modules of the tutorial by W3Schools (https://www.w3schools.com/python/default.asp) before the course (which requires approximately one day of work): Syntax, Variables, Data types, Operators, Lists, If…Else, While-Loops, For-Loops, Classes/Objects, and PIP.

Course aims

What this course is

This course …

(1) is a beginner's course for web scraping
(2) makes participants familiar with the problem of collecting massive data from Internet sources
(3) guides participants in evaluating the costs and benefits of automating data collection
(4) introduces participants to the structure of websites
(5) reviews the most effective approaches for collecting data from web sources
(6) provides hands-on implementations using Python
(7) outlines ethical and legal considerations

What this course is not

This course is not intended for experts in the domain of web scraping. It is primarily a beginner's course. We will start from scratch. We will discuss some advanced topics, but the focus is on providing a gentle introduction to web scraping for scientists.

Course objectives

Knowledge Objectives

- Participants are familiar with the problem of collecting massive amounts of data from the Internet
- Participants understand the structure of websites
- Participants understand the steps of the web scraping process
- Participants are aware of ethical and legal considerations

Skills Objectives

- Participants can evaluate the costs and benefits of developing web scrapers
- Participants can derive the requirements for a web scraper
- Participants can apply different web scraping techniques based on given problems

Learning Objectives

- Participants can program their own web scraper with Python
- Participants can apply the discussed methods to different websites
- Participants can apply advanced web scraping methods to complex problems

Preliminary schedule

The course will be held completely online via Zoom. Login details will be distributed after registration.

12.01.2023, 08:30-17:30: Day 1
13.01.2023, 08:30-17:30: Day 2
26.01.2022: Q&A (9:00-13.30)

Core readings

Programming

Chapagain, A. (2019). Hands-On Web Scraping With Python: Perform Advanced Scraping Operations Using Various Python Libraries And Tools Such As Selenium, Regex, And Others. Packt Publishing Ltd.

Mitchell, R. (2018). Web Scraping With Python: Collecting More Data From The Modern Web. O'Reilly Media, Inc.

Applications

Foerderer, J., Lueker, N., Heinzl, A. (2021): And the Winner is ...? Platform Governance, Awards, and Complementors' Product Strategies. Information Systems Research, 32(4).

Foerderer, J. (2020): Interfirm Exchange and Innovation in Platform Ecosystems: Evidence from Apple's

Worldwide Developers Conference, 66(10).

Foerderer, J., Kude, T., Mithas, S., Heinzl, A. (2018): Does Platform Owner's Entry Crowd Out Innovation? Evidence from Google Photos.

Course procedures

The course will be a mix of lectures and exercises. In the morning, the instructor will provide input. In the afternoons, participants work in teams on the course project (see below for details). We will have frequent Q&As to clarify questions regarding the course project.

Assessment

Form of assessment: course project (100%)

Participants will be assigned to work on a course project in teams. The course project is to develop a web scraper that collects data from an online shop. The course project has to be completed and submitted in the weeks after the course introduction.

Participation in all sessions is mandatory.

Workload

4 ECTS (22.5 hours lectures, 120 hours total workload)