JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

GGS Gießener Graduiertenzentrum Sozial-, Wirtschafts- und Rechtswissenschaften

Workshop

Web Data Science: Scraping and Analyzing the Web with Python

Organizational details

Instructor:	Dr. Jan Kinne
Dates:	1. Lecture: **October 4, 2021, 9:00 – 13:00**. Optional Q&A: 16:00 – 17:00. 2. Lecture: **October 11, 2021, 9:00 – 13:00**. Optional Q&A: 16:00 – 17:00. 3. Lecture: **October 18, 2021, 9:00 – 13:00**. Optional Q&A: 16:00 – 17:00. 4. Lecture: **October 25, 2021, 9:00 – 13:00**. Optional Q&A: 16:00 – 17:00. 5. Lecture: **November 1, 2021, 9:00 – 13:00**. Optional Q&A: 16:00 – 17:00. 6. Final Presentations: **November 8, 2021, 9:00 – 13:00**.
Location:	Online (Service TBA)
ECTS:	6 (4 if participant joins from Day 3)
Max. number of participants:	20

Objectives

In this course, participants will first receive a basic introduction to Python as a data science tool. Specific topics that will be covered include: Mapping, Web Scraping, and Text Mining. The course will conclude with the presentation of individual data analysis projects, which will be implemented by the participants in parallel with the course based on what they have learned. The goal is that by the end of the course, participants will have a basic understanding of Web Data Science and an overview of relevant tools, based on which one can develop further individually.

Content & Methods

- Day 1: Introduction to Python I (optional for advanced Python users)
  - Course overview
  - Introduction to Jupyter Notebooks
  - Python datatypes
  - Pandas dataframe basics
  - Pandas file I/O
  - Pandas dataframe functions I
  - Descriptive statistics with Pandas
  - Plotting with Pandas
- Day 2: Introduction to Python II (optional for advanced Python users)
  - Pandas dataframe functions II
  - Plotting using Seaborn
- Day 3: Mapping
  - Introduction to Geopandas
  - Geodata datatypes and I/O
  - Creating choropleth maps with Geopandas
  - Combining geodata and non-geographic data
  - Geocoding
- Day 4: Web Scraping
  - Basic introduction to HTML
  - Introduction to BeautifulSoup and Scrapy
  - Web crawling and web scraping
- Day 5: Text Mining
  - Labeled and unlabeled data for Machine Learning
  - Text preprocessing
  - Text vectorization
  - Train and test sets
  - Training a Machine Learning model
  - Calculating text similarities
- Day 6: Project Presentation
  - Participants present their own Data Science projects

The course will be held completely online. Each day there will first be a short introductory presentation by the lecturer. Afterwards, we will work together on a Data Science workflow using a pre-built and fully commented Jupyter notebook. These notebooks are prepared in such a way that they can be worked through again by the participants on their own after the course and parts of the workflow can be used for their own Data Science project. The participants will start such a project based on a self selected dataset on Day 1. Over the progression of the course, participants will explore their chosen dataset using the methods learned. Grades for the course are based on projects presented and submitted on the final day. In addition to the actual course days, there will be an additional meeting each week where participants can ask their individual questions concerning their projects.

Target group & Course Language

Target group: Everybody with an interest but little to no skills in web data and data science

Participation requirements: Installation of Python using the Anaconda installers for your OS:

- Windows: https://repo.anaconda.com/archive/Anaconda3-2020.11-Windows-x86_64.exe
- MacOS: https://repo.anaconda.com/archive/Anaconda3-2020.11-MacOSX-x86_64.pkg
- Linux: https://repo.anaconda.com/archive/Anaconda3-2020.11-Linux-x86_64.sh

Course language: English

Requirements for ECTS credits

- Active participation in the lectures of the course.
- At the beginning of the course, participants pick a dataset they are interested in from the Open Data website kaggle.com. Over the duration of the course, participants apply what they have learned to their chosen dataset, explore it on their own, and analyze it. In the final session, participants present the workflow they have created using their Jupyter notebooks. Grading is based on this notebook, which summarizes the entire Exploratory Data Analysis and its results.

About the instructor

Jan Kinne has been a member of the ZEW Research Department "Economics of Innovation and Industrial Dynamics" since June 2016. He studied geoinformatics at Heidelberg University and Loughborough University, UK. In the context of his dissertation (Dr. rer. nat.) at the University of Salzburg, which he successfully completed in September 2020, he developed web-based innovation indicators for microgeographic analyses. He continues to conduct method-oriented research on the use of (text-based) web data for innovation research. In 2019 he co-founded istari.ai, a startup that specializes in web-based company and market information in real-time.

Registration

By **September 24, 2021** via e-mail at info@ggs.uni-giessen.de.