
SPRING SCHOOL

# Webscraping und Data Mining mit Python

**14.03.–18.03.2022**

Johannes Gutenberg-Universität Mainz

**Referent: Dr. Jan R. Riebling**
**Lehrstuhl für Allgemeine Soziologie, Bergische Universität Wuppertal**

# SPRING SCHOOL

# Webscraping und Data Mining mit Python

## JGU Mainz, 14.03.-18.03.2022

**Referent:**
Dr. Jan R. Riebling
Lehrstuhl für Allgemeine Soziologie, Bergische Universität Wuppertal

**Inhalt:**
Durch die Allgegenwart von Informationstechnologien ist auch in den Sozialwissenschaften die Bedeutung digitaler Prozessdaten enorm gewachsen. Die adäquate Analyse solcher aus der Interaktion mit sozio-technischen Systemen gewonnen Daten erfordert entsprechendes Hintergrundwissen sowie spezifische praktische Fähigkeiten, die dieser Kurs vermitteln möchte. Durch Übungen und Anwendungsbeispiele sollen Möglichkeiten, aber auch Probleme neuer digitaler Datenquellen aufgezeigt werden. Der Workshop führt zunächst in die Programmiersprache Python ein. Danach werden verschiedene Methoden der Datenextraktion und schließlich auch die speziellen Herausforderungen der Aufbereitung und Analyse dieser Datentypen behandelt.

**Veranstaltungsformat:**
Die Spring School findet in deutscher Sprache statt. Aktuell ist die Veranstaltung in Präsenz geplant. Gegebenenfalls muss pandemiebedingt kurzfristig auf ein Online-Format gewechselt werden.

**Teilnahmegebühr und Anmeldung:**
JGU-Mitglieder: 180 €
Externe Teilnehmende: 500 €

Link zur Anmeldung:
sozialstruktur.soziologie.uni-mainz.de/springschool2022

Organisation:
Arbeitsbereich Soziologie und quantitative Methoden
Prof. Dr. Natascha Nisic
Arbeitsbereich Sozialstrukturanalyse
Prof. Dr. Gunnar Otte
Institut für Soziologie
Johannes Gutenberg-Universität Mainz
Kontakt: springschool2022@uni-mainz.de

JOHANNES GUTENBERG UNIVERSITÄT MAINZ JGU

Interdisciplinary Public Policy Mainz IPP

**Programm:**
Die Spring School ist auf 5 Tage ausgelegt und besteht etwa zur Hälfte aus Vortrags- und Übungseinheiten. Die Abfolge der Themen wird flexibel gehalten und kann an die Bedürfnisse und Forschungsfragen der Teilnehmenden angepasst werden.

| | |
|---|---|
| **Tag 1 – 14.03.2022** | **Computational Social Science und Einführung in Python I**<br>Einführung in die technische Infrastruktur und begleitete Installation benötigter Programmpakete sowie die Vorstellung grundlegender Programmiertechniken und Vorgehensweisen im Rahmen der Computational Social Science (CSS). Im Rahmen des Workshops wird hauptsächlich die Anaconda Distribution verwendet werden. |
| ab 12:00 | Anmeldung |
| 13:00–18:00 | Das Medium Data Problem in CSS, Python: Anaconda Distribution, Jupyter Notebook and JupyterLab als grundlegende Infrastruktur |
| **Tag 2 – 15.03.2022** | **Einführung in Pyton II**<br>Ausführliche Besprechung von Pythons Syntax, primitiven und weiterführenden Datentypen sowie relevanter Python-Module und Bibliotheken für die Forschung. |
| 09:00–12:30 | Grundlegende Syntax von Python, Datentypen und Typenhierarchie, Container-Objekte, Schleifen und logische Bedingungen |
| 12:30–14:00 | Mittagspause |
| 14:00–18:00 | Funktionen und Klassen, Data Frames und Serien, Scientific Computing |
| **Tag 3 – 16.03.2022** | **APIs und Webscraping**<br>Umgang mit webbasierten APIs (Application Programming Interfaces) im Zuge der Datenerhebung. Generelles Vorgehen beim Webscraping, dem rekursiven Download von Webressourcen und der Extraktion von Informationen aus HTML-Dokumenten. |
| 09:00–12:30 | URLs als Basis von Client-Server-Interaktion, Einsatz von Web-APIs, Webcrawling |
| 12:30–14:00 | Mittagspause |
| 14:00–18:00 | HTML-Dokumente, Webspcraping vs. Webcrawling |
| ab 19:00 | Gemeinsames Abendessen |
| **Tag 4 – 17.03.2022** | **Datenaufbereitung**<br>Aufbereitung von Daten und die Extraktion von Informationen aus Rohdaten, wie z.B. Texten. Als zentrale Technik wird hier die Arbeit mit regulären Ausdrücken vorgestellt und eingeübt. |
| 09:00–12:30 | Reguläre Ausdrücke, Text Mining |
| 12:30–14:00 | Mittagspause |
| 14:00–18:00 | Data Munging: Von den Rohdaten zum DataFrame, Webdriving und Spoofing |

| | |
|---|---|
| **Tag 5 - 18.03.2022** | **Datenmanagement und Datenanalyse**<br>Techniken des Umgangs mit Daten und deren Archivierung. Wie ist die Datenqualität von prozessgenerierten Daten zu beurteilen? Ein weiteres Thema ist die Vorstellung von Analysetechniken, wie beispielsweise dem Machine Learning, und der Verweis auf weiterführende Ressourcen zum Selbststudium. |
| 09:00-12:00 | Datenpersistenz: Datenbanken und Datenformate, Fragen der Datenqualität, Statistische Modellierung in Python |
| 12:00-13:00 | Mittagspause |
| 13:00-15:00 | Machine Learning |
| 15:00 | Ende der Spring School |