Praxisorientierte Einführung in Data Scraping

Inhaltliche Beschreibung:

Mit zunehmender Verfügbarkeit digitaler Daten wird das Schlagwort "Big Data" häufig verwendet, um Informationen über Online-Umgebungen wie Webseiten und soziale Netzwerke zu beschreiben. Während solche Daten für eine Vielzahl von Forschungsgebieten relevant sind, ist das Abrufen und Verarbeiten – Data Scraping - für Sozialwissenschaftler häufig eine methodische Hürde. Dieser Kurs bietet eine praktische und systematische Einführung in die Beschaffung solcher Daten mit der Open-Source-Programmiersprache Python ab. Die TeilnehmerInnen lernen den Umgang mit automatisierten Methoden zum Abrufen von Daten aus Programmierschnitstellen (APIs) wie Twitter, sowie von Webseiten und PDF-Dateien. Nach erfolgreichem Abschluss dieses Kurses können die Teilnehmer/-innen selbstständig Data Scraping Projekte für sozialwissenschaftliche Forschung durchführen.

Teilnahmevoraussetzungen:

Teilnehmer/-innen sollten bereits Erfahrung im Umgang mit einer syntaxbasierten Statistiksoftware (z.B. Stata) oder Programmiersprache (z.B. R) haben. Zudem sollten grundlegende Kenntnisse im Bereich der deskriptiven und induktiven Statistik vorhanden. Es wird zudem empfohlen, dass Teilnehmer/-innen ihre eigenen Rechner (z.B. Notebooks) mitbringen und für die Bearbeitung der Kursinhalte verwenden.

Leistungsnachweise:

Um den Kurs erfolgreich zu bestehen, müssen TeilnehmerInnen zwei Portfolio Leistungen erbringen, die das Wissen über die Programmiersprache Python und die Datenbeschaffung über Programmierschnittstellen und Webseiten prüfen. Im Zuge dieser Leistungsnachweise sollen mehrere Programmieraufgaben sowie ein Abschlussprojekt bearbeitet werden.

Literatur:

- 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..