PCMO@GGS-Workshop

Datensammlungen mit APIs: Die Plattform Wikipedia als Beispiel

Organisatorisches

Kursleitung:	Felix Soldner & Leon Fröhling
Termin:	13. Oktober 2023, 9.00 – 17.00 Uhr
Ort:	Erwin-Stein-Gebäude, Goethestraße 58, 35390 Gießen, Raum 201 (Großes Sitzungszimmer)
Max. Anzahl Teilnehmende:	12

Veranstaltungsziel

Teilnehmende lernen in diesem Kurs:

- wie APIs aufgebaut sind und genutzt werden können, um verschiedene Arten von Web-Daten zu sammeln.
- wie über eine API verschiedene Ressourcen von Wikipedia automatisiert gesammelt werden können.
- wie gesammelte Web-Daten für weitere Analysen eingelesen, aufbereitet und wieder gespeichert werden können.
- wie kritisch über automatisiert gesammelte Web-Daten reflektiert werden kann, um mögliche systematische Verzerrungen zu identifizieren und zu dokumentieren.

Inhalt & Methode

Der Workshop ist darauf ausgelegt, in komprimierter Form die Grundlagen des Sammelns und der Verarbeitung von Web-Daten zu vermitteln. Er beschränkt sich dabei auf die Arbeit mit APIs (*Application Programming Interfaces*), einer weit verbreiteten und in der grundlegenden Funktionsweise weitestgehend universellen Art der Verfügbarmachung von Daten. Vorgestellt wird die typische Interaktion mit einer API am Beispiel von Wikipedia, eine Plattform die verschiedenste Ressourcen über einen API-Zugang zur Verfügung stellt.

Der Workshop ist gegliedert in kurze theoretische Hinführungen zu verschiedenen Aspekten des Arbeitens mit Web-Daten, die stets von praktischen “Live-Coding”-Blöcken, in denen Möglichkeiten der automatisierten Interaktion mit APIs und den gesammelten Daten vorgestellt werden, begleitet sind. Zusätzlich ist ausreichend Zeit für das eigenständige Ausprobieren der vorgestellten Methoden und Verarbeitungsschritte durch die Teilnehmenden eingeplant.

Der grobe inhaltliche Aufbau des Workshops ist wie folgt:

1. Einführung zum Sammeln von Web-Daten
2. Einführung zum Arbeiten mit APIs und erstes Ausprobieren der Wikipedia API im Browser
3. Programmierung von API Interaktionen zur automatisierten Sammlung von Daten
4. NLP Grundlagen und Verarbeitung gesammelter Daten
5. Wissenschaftliche Best-Practices für den Umgang mit Web-Daten

Die vorgestellten Programmierungen und automatisierten Interaktionen werden anhand der Programmiersprache Python gezeigt. Um insbesondere in den eingeplanten Übungen eigenständig Fortschritte und Erfolge zu erzielen sind Vorkenntnisse der Grundlagen der Programmierung im Allgemeinen und die Syntax von Python im speziellen wünschenswert (z.B. die allgemeine Funktionsweise und Implementierung von einfachen *for-loops* und *if-else-conditions*). Die Dozenten werden während der Übungen zusätzlich mit Rat und Tat zur Verfügung stehen, um die Teilnehmenden bei kleineren Schwierigkeiten der Umsetzung zu unterstützen. Die Übungen für die Teilnehmenden werden in Google Colab vorbereitet und durchgeführt, um eine einheitliche Programmierungsumgebung zu schaffen. Um Google Colab nutzen zu können, wird ein Google Konto benötigt.

Zielgruppe & Veranstaltungssprache

Der Workshop richtet sich an Postdocs und fortgeschrittene Promovierende des GGS.

Veranstaltungssprache: Deutsch

Teilnahmevoraussetzung:

- Google Account (um die Nutzung von Google Colab zu ermöglichen)
- Persönlicher Computer (Laptop)

Über die Referenten

Felix Soldner ist wissenschaftlicher Mitarbeiter in der Abteilung Computational Social Science in GESIS – Leibniz-Institut für Sozialwissenschaften mit einem Hintergrund in Psychologie, Kognitionswissenschaften und Kriminalwissenschaften. Seine Forschung umfasst Themen im Zusammenhang mit Online-Betrug, Täuschungserkennung, Kryptomärkten und Fehler in Datenerhebungen. Für seine Forschung sammelt er Daten von verschiedenen Online-Plattformen und nutzt maschinellen Lernens sowie natural language processing für Analysen von Texten.

Leon Fröhling ist wissenschaftlicher Mitarbeiter in der Computational Social Science Abteilung der GESIS, und arbeitet parallel an seiner Dissertation an der RWTH Aachen. Seine Forschung beschäftigt sich mit der Messung, Gewährleistung und Verbesserung der Qualität neuartiger Datentypen im Bereich der Computational Social Science. Hauptgegenstand sind dabei die

systematische Untersuchung der Auswirkungen von Design-Entscheidungen des Prozesses der Datensammlung und –verarbeitung auf resultierende Datensätze und Forschungsergebnisse.

Anmeldung

Wenn Sie an der Veranstaltung teilnehmen möchten, melden Sie sich bitte bis zum **3. Oktober 2023** per E-Mail an postdocs@ggs.uni-giessen.de an.