Titel: Web-Scraping in R

Datum: 15.02.2021

Uhrzeit: 10:00 – 15:00 Uhr

Format: Online-Session

Beschreibung:

Content

In diesem Workshop wird ein Einblick in wichtige Schritte und Verfahren zur Verfügbarmachung von Daten aus Online-Quellen in R gegeben. Häufig stehen interessierende Daten zu verschiedensten Themen nicht in vorgefertigten Datensätze zum Download bereit. Beispielhaft wäre die Berichterstattung von Zeitungen bzgl. rassistischer Gewalt in den letzten Jahren oder die Erfassung lobbyistischer Aktivitäten im dt. Bundestag/Ministerien zu denen Informationen auf tausenden Internetseiten der Verwaltung veröffentlicht wurden. Manuelles Kopieren relevanter Informationen ist in vielen dieser Szenarien unpraktikabel.

Eine Lösung bietet die automatisierte Auslesung relevanter Informationen aus dem Web, bei gleichzeitigem Aufbau eines Datensatzes mittels eines Statistikprogramms. R bietet verschiedene Möglichkeiten diese Aufgaben zu erfüllen.

In dem Workshop werden dazu folgende Themenkomplexe vorgestellt:

* Grundstruktur von Web-Mining (Web-Crawling, Web-Scraping, Data-Parsing)
* Grundaufbau von Websites (html, xml, css)
* Lokalisierung interessierender Daten auf Websites (css- & xpath)
* Dynamische vs. Statische Websites
* Einfaches Web-Scraping statischer Websites (rvest)
* Web-Scraping dynamische Websites mittels automatisierter Web-Browser Steuerung (RSelenium)
* Datenaufbereitung

WICHTIGER HINWEIS

Der Workshop wird in Form einer Zoom-Session durchgeführt. Sie müssen selbst keine Webcam besitzen oder diese aktivieren - es ist lediglich nötig, dass Sie die Möglichkeit haben, Töne auf Ihrem Computer auszugeben. Außerdem benötigen Sie eine aktuelle Version von R und RStudio.

ZIELGRUPPE

Studierende und Promovierende mit Grundkenntnissen in R, die Erhebungen von Daten aus Onlinequellen durchführen möchten.