

SSRPCB-2: Web Scraping with Python and LLMs for Social Work Research: Methods, Ethics, and Legal Considerations

Thursday, January 21, 2027, 8:00 am – 10:00 am

Presenters:

Nari Yoo, PhD (University of Michigan) and Cheng Ren, PhD (University at Albany)

Description:

Web-based data offers social work researchers information that traditional sources cannot provide. Provider directories, government databases, online forums, and organizational websites contain large volumes of publicly available data that can be collected non-intrusively, at scale, and at low cost. For example, participants might scrape state licensing board directories to study the geographic distribution of licensed social workers and therapists, federal directories such as the Office of Refugee Resettlement (ORR) state resource map to identify refugee resettlement agencies, or community service directories such as Michigan 211 to map local availability of housing, food, and behavioral health resources. This two-hour workshop teaches participants how to collect web-based data using Python, covering the core libraries BeautifulSoup and Selenium, and demonstrates how large language models (LLMs) can extend traditional scraping by parsing unstructured pages, extracting targeted information through prompts, and producing structured outputs from messy HTML. The workshop also covers legal and ethical issues: compliance with website terms of service, the Computer Fraud and Abuse Act (CFAA), IRB considerations when scraping publicly available but potentially sensitive data, and responsible storage and use of scraped content. Participants will leave with practical experience using Python and LLMs to collect web data, annotated Jupyter notebooks they can adapt for their own projects, and a clear sense of when a given site can be scraped legally and ethically. The workshop uses live coding on social-work-relevant sites, walk-throughs of shared code, and a short hands-on exercise with instructor support. The workshop is designed for doctoral students, postdoctoral fellows, and faculty interested in collecting non-survey web data; basic Python familiarity (variables, functions, working with pandas DataFrames) is recommended but not required, though participants without prior Python experience can still benefit from the methods overview and the legal/ethical content.

Agenda:

Outline with times and topics/activities/objectives of how the event will run

0:00 - 0:10 Welcome, introductions, and motivation: when is web scraping the right tool for social work research?

0:10 - 0:30 HTML/HTTP foundations and core scraping with BeautifulSoup; live demo on a sample provider directory.

0:30 - 0:50 Handling dynamic, JavaScript-rendered pages with Selenium; pagination, waits, and common failure modes.

0:50 - 1:15 LLM-assisted extraction: prompting strategies for parsing messy HTML, producing structured outputs, and validating LLM-extracted fields.

1:15 - 1:40 Legal and ethical considerations: terms of service, the Computer Fraud and Abuse Act (CFAA), Fair Use, IRB review of publicly available but potentially sensitive data, and secure storage practices.

1:40 - 1:55 Hands-on guided exercise and walkthrough ("should I scrape this site?").

1:55 - 2:00 Q&A, shared resources, and next steps.