DRDHum 2024 Workshop proposal

FIN-CLARIAH tools to make sense of web data

Organizers: Inés Matres (ines.matres@helsinki.fi, University of Helsinki), Mietta Lennes (mietta.lennes@helsinki.fi, kielipankki.fi), Masoud Fatemi (masoud.fatemi@uef.fi, University of Eastern Finland)

This 2-hour workshop presents the results, services, and ongoing work produced within FIN-CLARIAH (https://www.kielipankki.fi/organization/fin-clariah/). This research infrastructure is currently funded by the Research Council of Finland, and its activities aim at fostering data-intensive and digital research in Social Sciences and Humanities (SSH). The workshop introduces datasets and tools that are freely available for SSH researchers, including newspapers, periodicals, and other publications from the National Library of Finland, machine-readable records from the National Archives, Finnish parliamentary speeches, Twitch game streams, and social media data from the Nordic region. These novel tools and interfaces have been built to evaluate, subset, enrich, and analyze large-scale SSH datasets. Current efforts are directed towards supporting visual and multimodal research and developing transformer models for SSH research. In addition to introducing resources available, in this workshop we will focus on a selection of tools and datasets for social media and web data.

Overall, the FIN-CLARIAH consortium comprises two components, FIN-CLARIN and DARIAH-FI. The Language Bank of Finland (Kielipankki) provides centralized services for sharing and reusing materials and tools in the research community. In turn, the DARIAH-FI consortium consists of SSH research teams with high demands and expertise in data-intensive research committed to make what they develop (datasets, tools and methods) available to wider research communities.

This workshop offers an experimental and hands-on setting that complements the conference theme on digital applications in the advent of ML and AI. After providing an overview of FIN-CLARIAH and its core services, there will be a practical section with four resources for researchers in the format of a brief tutorial with time for attendees to try the showcased resources on their own laptops and to pose questions to the presenters. The workshop is open to all conference participants.

Preliminary schedule:

- Introduction to FIN-CLARIAH resources for SSH research, 25 minutes
- Brief tutorial on Nordic Tweet Stream (NTS), a multilingual monitor corpus of geolocated tweets and associated metadata from the Nordic region covering the period 2013-2023. The data was collected using the academic API which is now closed. (20 minutes)
- Brief tutorial on TurkuNLP tools, machine learning tools to annotate and identify toxic language, genre and interaction in web content (20 minutes)

- Brief tutorial on subsetting data from social media, participants will learn how to explore large datasets that have not originally been created for research and extract the subset they are interested in. (20 minutes)
- Brief tutorial on services in the Language Bank of Finland for research on social media data, 20 mins
- General discussion: 15 min

Keywords: social media, datasets, subsetting data, analysis tools