The COVID-19 Open Research Dataset

by Lucy Lu Wang and Kyle Lo

Abstract: The COVID-19 Open Research Dataset (CORD-19) is a growing corpus of scientific articles on COVID-19 and related historical research on other coronavirus outbreaks and epidemics. Since its release on March 13th, 2020, CORD-19 has grown to include over 45K articles (of which over 30K have machine-readable structured full text) and has been widely adopted for various text mining and information retrieval applications. In this talk, we discuss the curation of CORD-19 and promising avenues of research conducted over the corpus. We hope this resource will bring together the computing community, clinical experts, and policy makers in the search for effective treatments and management policies for COVID-19.

Biographies: Lucy Lu Wang is a Young Investigator at the Allen Institute for AI, where she is part of the Semantic Scholar Research team. She works on increasing access to and understanding of scientific text, with a focus on applications in bioNLP, biomedical ontology, and the science of science. Her work on gender trends in publishing and supplement interaction detection has been featured in Geekwire, Gizmodo, Axios, VentureBeat, and the New York Times. She completed her PhD at the University of Washington in Biomedical and Health Informatics.
Kyle Lo is a research scientist at the Allen Institute for AI on the Semantic Scholar Research team, where he works on NLP for scientific text with emphasis on knowledge extraction, summarization, and discourse. His recent work has been in domain adaptation of language models for improved scientific text mining and AI-assistive tools for improving research workflows. His work on programatically identifying sex bias in clinical trial participation published in JAMA was featured in Quartz. He has an MS in Statistics from the University of Washington.