Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset

Submitted by Anonymous on Wed, 03/18/2020 - 8:35am

THE WHITE HOUSE
Office of Science and Technology Policy

FOR IMMEDIATE RELEASE
March 16, 2020

Today, researchers and leaders from the Allen Institute for AI, Chan Zuckerberg Initiative (CZI), Georgetown University's Center for Security and Emerging Technology (CSET), Microsoft, and the National Library of Medicine (NLM) at the National Institutes of Health released the COVID-19 Open Research Dataset (CORD-19) of scholarly literature about COVID-19, SARS-CoV-2, and the coronavirus group.

Requested by The White House Office of Science and Technology Policy, the dataset represents the most extensive machine-readable coronavirus literature collection available for data and text mining to date, with over 29,000 articles, more than 13,000 of which have full text.

Now, The White House joins these institutions in issuing a call to action to the Nation's artificial intelligence experts to develop new text and data mining techniques that can help the science community answer high-priority scientific questions related to COVID-19.

The collection was constructed via a unique collaboration between Microsoft, NLM, CZI, and the Allen Institute for AI, coordinated by Georgetown University. Microsoft's web-scale literature curation tools were used to identify and bring together worldwide scientific efforts and results, CZI provided access to pre-publication content, NLM provided access to literature content, and the Allen AI team transformed the content into machine-readable form, making the corpus ready for analysis and study. The CORD-19 resource is available on the Allen Institute's SemanticScholar.org website https://pages.semanticscholar.org/coronavirus-research and will continue to be updated as new research is published in archival services and peer-reviewed publications. Researchers should submit the text and data mining tools and insights they develop in response to this call to action via the Kaggle platform https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge . Through Kaggle, a machine learning and data science community owned by Google Cloud, these tools will be openly available for researchers around the

world.

To inform the call to action, key scientific questions related to COVID-19 were developed in coordination with the National Academies of Sciences, Engineering, and Medicine's Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats and the World Health Organization.

The call to action and key questions are both available on Kaggle. "Decisive action from America's science and technology enterprise is critical to prevent, detect, treat, and develop solutions to COVID-19. The White House will continue to be a strong partner in this all hands-on-deck approach. We thank each institution for voluntarily lending its expertise and innovation to this collaborative effort, and call on the United States research community to put artificial intelligence technologies to work in answering key scientific questions about the novel coronavirus," said Michael Kratsios, U.S. Chief Technology Officer, The White House.

"This valuable new resource is the fruit of unselfish collaboration and now offers the opportunity to find answers to important questions about COVID-19," said Dr. Dewey Murdick, Director of Data Science at Georgetown University's Center for Security and Emerging Technology (CSET), who coordinated the cross-team effort. "Once the crisis has passed, we hope this project will inspire new ways to use machine learning to advance scientific research."

"It's all-hands on deck as we face the COVID-19 pandemic," said Dr. Eric Horvitz, Chief Scientific Officer at Microsoft. "We need to come together as companies, governments, and scientists and work to bring our best technologies to bear across biomedicine, epidemiology, AI, and other sciences. The COVID-19 literature resource and challenge will stimulate efforts that can accelerate the path to solutions on COVID-19." "Sharing vital information across scientific and medical communities is key to accelerating our ability to respond to the coronavirus pandemic," said Dr. Cori Bargmann, Head of Science at the Chan Zuckerberg Initiative. "The new COVID-19 Open Research Dataset will help researchers worldwide to access important information faster."

"We are excited to be part of this collaboration to aid in the COVID-19 response and that the group is making use of our open access subset on coronavirus literature," said Dr. Patricia Flatley Brennan, Director of the National Library of Medicine at the National Institutes of Health. "Our current collection of more than 10,000 full-text scholarly articles related to coronavirus provides a critical resource for text mining efforts like this one."

"One of the most immediate and impactful applications of AI is in the ability to help scientists, academics, and technologists find the right information in a sea of scientific papers to move research faster. We applaud the OSTP, WHO, NIH and all organizations that are taking a proactive approach to use the most advanced technology in the fight against COVID-19," said Dr. Oren Etzioni, Chief Executive Officer of the Allen Institute for AI. "The Allen Institute for AI, and particularly the Semantic Scholar team, is committed to updating and improving this important

resource and the associated AI methods the community will be using to tackle this crucial problem."

"It's difficult for people to manually go through more than 20,000 articles and synthesize their findings. Recent advances in technology can be helpful here. We're putting machine readable versions of these articles in front of our community of more than 4 million data scientists. Our hope is that AI can be used to help find answers to a key set of questions about COVID-19," said Anthony Goldbloom, Co-Founder and Chief Executive Officer at Kaggle. For more information about the novel coronavirus and COVID-19, please visit: https://www.cdc.gov/coronavirus

LINK
https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/