Machine Learning and Digital Classical Chinese Texts: Collaboration between the UC Computing Platform and Peking University's Big-Data databases.

Hu, Minghui
mhu@ucsc.edu
University of California Santa Cruz, United States of America

Li, Xiao
xli422@ucsc.edu
University of California Santa Cruz, United States of America

Weekley, Jeffrey
jweekley@ucsc.edu
University of California Santa Cruz, United States of America

We report an ongoing effort to apply state-of-the-art machine learning techniques on the massive and independent accumulation of classical Chinese databases, as a repeatable and malleable prototype using a continuous integration/continuous development (CI/CD) framework, interactive Jupyter Notebooks, shared data, and open-source software on the Nautilus Hyper-converged, distributed cluster. Nautilus is a cloud-like, national-level GPU/CPU resource; part of the NSF-supported National Research Platform; and has participants in North America, the Pacific Rim, and Europe. Nautilus allows researchers to scale from their laptop to hundreds of graphics processors or thousands of CPU cores easily and flexibly. Rather than recreating existing, massive Chinese databases (with their inherent and competing commercial interests and copyright issues), we leverage two databases co-developed by the National Library of China and the Research Center for Digital Humanities of Peking University to develop our prototype approaches. In order to avoid complex layers of international negotiation and agreement of replicating databases, we use the adaptability and flexibility of Nautilus to craft data-agnostic reproducible approaches and machine-learning toolkits that are generalizable yet customizable for many research investigations. We will report on current progress and future plans for this project.

Bibliography

Altintas, I., et al. (2019, May). Workflow-driven distributed machine learning in CHASE-CI: A cognitive hardware and software ecosystem community infrastructure. In 2019 IEEE international parallel and distributed processing symposium workshops (IPDPSW).

Bleeker, E., et al. (2022). A Game of Persistence, Self-doubt, and Curiosity: Surveying Code Literacy in Digital Humanities. DH Benelux Journal.

Liu, A., et al (2017). Open, Shareable, Reproducible Workflows for the Digital Humanities: The Case of the 4Humanities.org" WhatEvery1Says" Project.

Long, L., et al (2022, May). Composable Infrastructures for an Academic Research Environment: Lessons Learned. In 2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW).

Oberbichler, S., et al (2022). Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. Journal of the Association for Information Science and Technology.

Olaru, O. (2019). What is Digital Humanities and What's It Doing in Romanian Departments?. Revista Transilvania.

Sherratt, T. (2019). GLAM workbench: Possibilities of collection data for research, experimentation, and collaboration.

Tan, C. R. (2021). The nascent case for adopting Jupyter notebooks as a pedagogical tool for interdisciplinary Humanities, Social Science, and Arts education. Computing in Science & Engineering.