Andrea Wen-Yi Wang

aww66@cornell.edu || https://andreawwenyi.github.io/

EDUCATION

PhD Information Science, Cornell University Aug 2022 - Present

Chairs: Allison Koenecke, David Mimno; **Committee**: Karen Levy

MS Data Science, New York University 2017 - 2019

Mathematics and Data track

BA Finance, National Taiwan University 2012 - 2016

RESEARCH STATEMENT

My research interests lie at the intersection of Natural Language Processing, Data Science, and Computational Social Science. My works involve understanding the characteristics of large language models (LLMs) in two ways. First, I study the interpretability of multilingual large language models with the goal to improve the performance for low-resource languages. Second, I study both the opportunities and challenges that LLMs offer for social scientists with textual data. I have works in domains related to criminal justice, gendered studies, misinformation.

PUBLICATIONS

1. **Andrea W Wen-Yi** and David Mimno. 2023. *Hyperpolyglot LLMs: Cross-Lingual Interpretability in Token Embeddings.* In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 1124–1131, Singapore. Association for Computational Linguistics.
2. **Andrea Wang**, Jo-Yu Lan, Ming-Hung Wang, Chihhao Yu. *The Evolution of Rumors on a Closed Social Networking Platform During COVID-19: Algorithm Development and Content Study.* JMIR Med Inform 2021;9(11):e30467. doi: 10.2196/30467

WORKING PAPERS

1. (Under Review) **Andrea W Wen-Yi**, Kathryn Adamson, Nathalie Greenfield, Rachel Goldberg, Sandra Babcock, David Mimno, Allison Koenecke *"Courtroom Tears": Identifying Gendered Discourse in US Capital Trial Transcripts Using Large Language Models.*
2. **Andrea W Wang**, Allison Koenecke, David Mimno. *Seasonality Visualizations of Online Text.*
   - Presented at *IC2S2 2023 (Poster)*

WORK/RESEARCH EXPERIENCES

Data Scientist ***Public Safety Lab, New York University*** Jan 2019 - June 2022

- Created and maintained a system collecting individual-level detainee records daily from around 1,000 U.S. county jail rosters in the US with Python requests, selenium, beautifulSoup.
- Designed operational pipeline involving Amazon Web Services, Github, and Airtable that improved the success rate by 80%.
- Built API with Node.js and web interfaces with a MEAN (MongoDB, Express, Angular, Node) stack.

Data Researcher ***Information Operations Research Group*** Sep 2020 - Aug 2021

- Developed an efficient clustering algorithm to identify messages that belong to the same rumor and studied the temporal propagation patterns of false pandemic-related rumors on LINE.

Core Software Developer ***g0v (gov-zero)*** Oct 2019 - Jul 2020

- Developed and managed the *0archive* project, an open source automated archiving system for Taiwanese news medias websites, forums, and content farms to provide open texts data for civic and academic researchers.

TEACHING

1. **INFO2950: Introduction to Data Science** *Cornell Information Science* Fall 2023
2. **Probability and Statistics for Data Science (Graduate-level)** *NYU Center for Data Science* Fall 2018
3. **Introduction to Data Science (Graduate-level)** *NYU Center for Data Science* Spring 2019

TALKS

1. COSCUP 2021. Information operations research as a data science research. Virtual. Jul 2021.
2. Chung Yuan Christian University Graduate Seminar. From NTU Finance to NYU Center for Data Science: My experiences and learnings. Taoyuan, Taiwan. May 2021.
3. g0v summit 2020. 0archive - an open source archiving system and open data for Taiwan online information space. Tainan, Taiwan. Dec 2020.

Skills

- **Programming Languages**: Python, SQL, R, Javascript, Matlab, Linux
- **Cloud Platform** AWS
- **Databases** MySQL, MongoDB, Google BigQuery