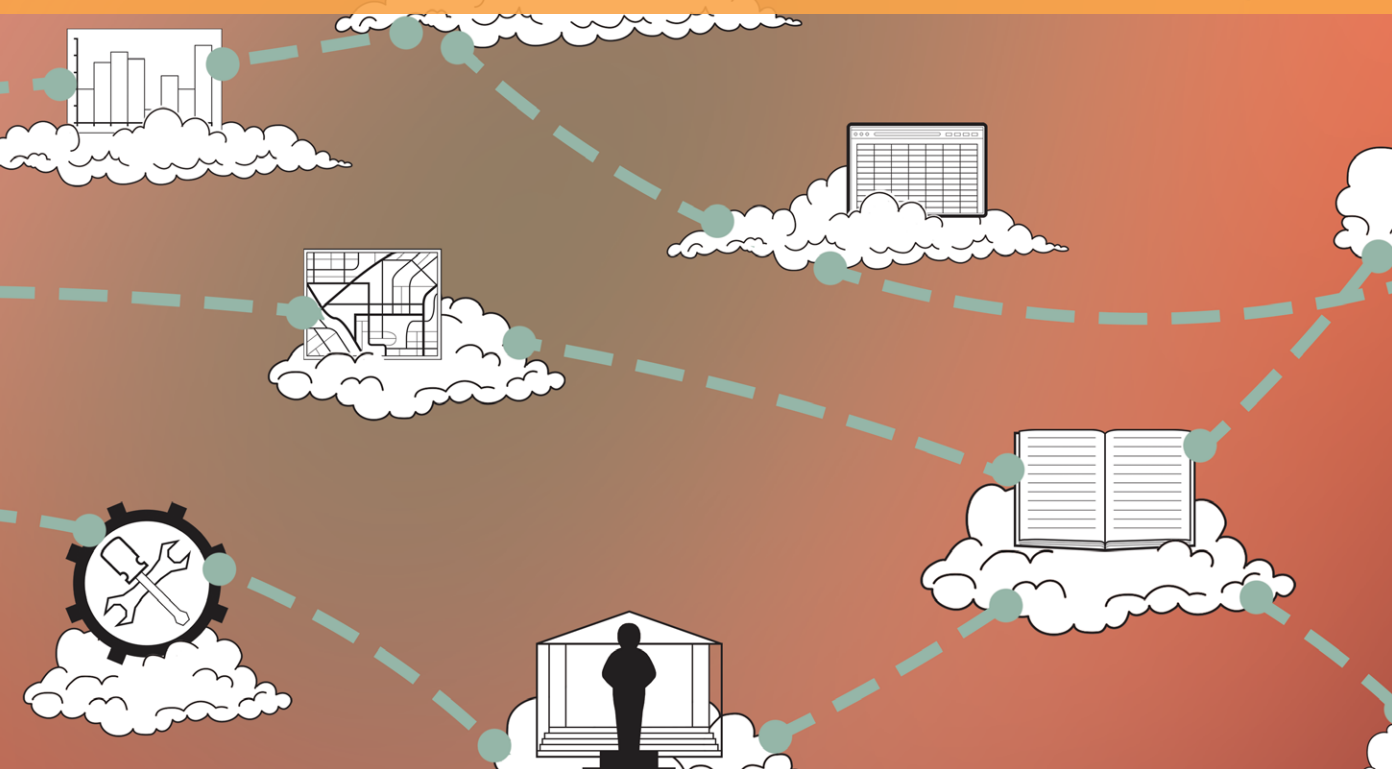
CESTA research anthology 2022

This is an excerpt from the *CESTA Research Anthology 2022*. To read the full *Anthology*, click here. To learn more about CESTA's *Anthology* series and read editions from previous years, click here.



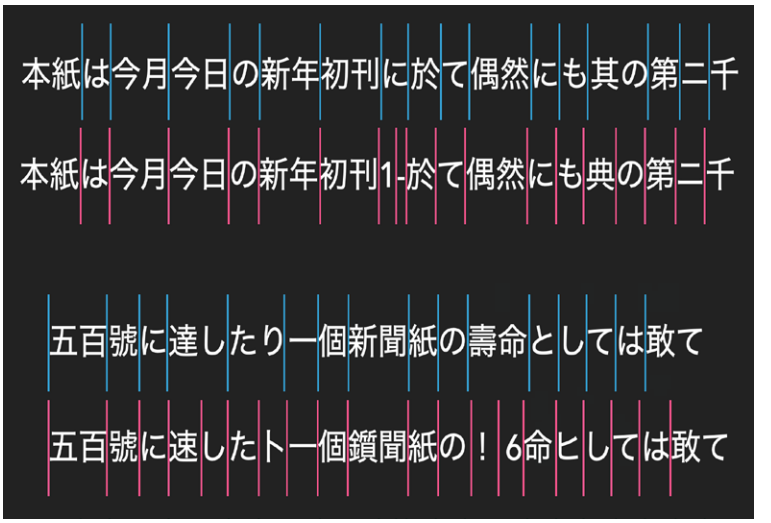
Imperial Vocabulary: Public Political Discourse of Trans-Pacific Japan, 1868-1912

by Andrew Nelson (East Asian Languages and Cultures)

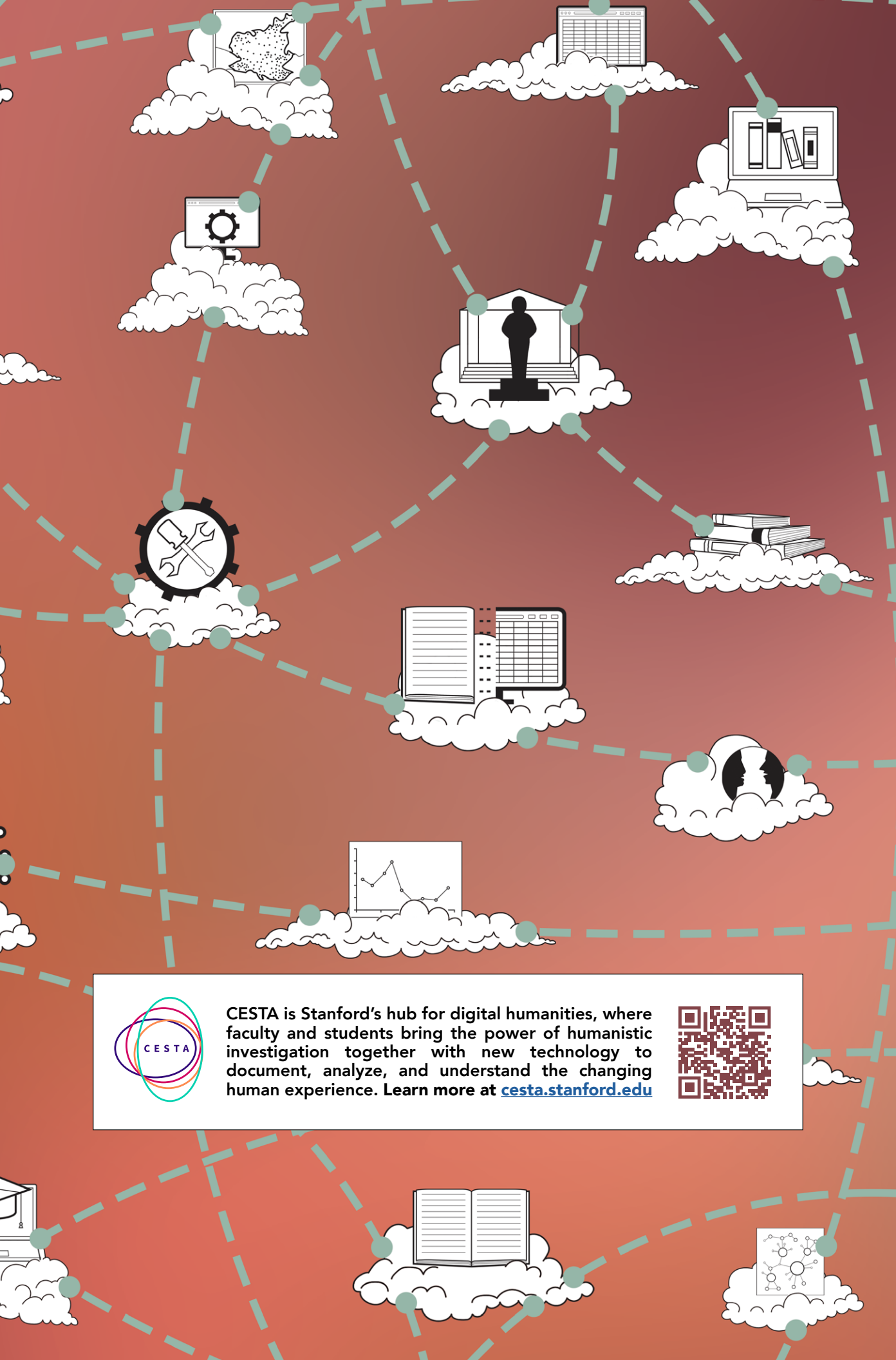
In this project, I developed a model for querying word frequencies in the Hoji Shimbun Japanese Diaspora newspaper collection in order to trace occurrences of politically-related terms over the period from 1895 to 1935. The project leverages a corpus of over 623 million words extracted from over 400,000 pages of text from 10 newspapers published in San Francisco, Beijing/Dalian, Honolulu, and São Paulo.

Evaluating the accuracy of existing optical character recognition (OCR) methods was a major component of the project, and an area where I benefited from the support of my interns, Victor Cheruiyot and Sandi Khine. Since political terms are written in kanji, we were particularly interested in kanji OCR accuracy. Hand-grading a sample of approximately 25,000 OCR output characters revealed a kanji accuracy rate of approximately 80% for the best-performing OCR method. I improved baseline accuracy by removing alphanumeric characters and Western punctuation, which were almost entirely absent from the original newspaper texts, and by removing all kana characters that had become isolated when erroneous characters fragmented real words during tokenization.

Querying political terms confirmed well-known historical narratives, and invited the investigation of new questions. For example, the frequency of "modern" (現代) increases through the 1920s in both the *Manshū Nichinichi* and the *Burajiru Nippō*, which were both aligned with the Japanese metropolitan government. Frequencies of "modern" in the San Francisco and Hawai'ian newspapers parallel each other but find no parallel in the Chinese and Brazilian papers, suggesting that discourses of modernity followed their own contours in American Japanese communities.



This image shows two lines of parsed text. The upper line in each pair, shown with blue parsing lines, is the actual newspaper text reproduced manually. The lower line, shown with pink parsing lines, is the corresponding OCR output. Word count differences following punctuation and stopword removal can serve as proxies for OCR accuracy.



CESTA is Stanford's hub for digital humanities, where faculty and students bring the power of humanistic investigation together with new technology to document, analyze, and understand the changing human experience. **Learn more at** cesta.stanford.edu