Tracing the shifting rhetoric of ethnoracial difference in federal responses to education, 1958-2018

Project investigator: Andrés Castro Samayoa

The U.S. Federal Documents Collection

HathiTrust's U.S. Federal Documents Collection offers a unique opportunity for researchers to explore a trove of digitized documents that can help us deepen our historical understanding of shifts in federal initiatives. For researchers interested in how the United States has addressed issues of racial inequality across multiple federal agencies, the U.S. Fed Docs collection offers a novel trove of data for large-scale textual analyses. can serve as a promising starting point.

Creating a Dataset

In this ACS project, Andrés Castro Samayoa (assistant professor at Boston College) has worked closely alongside HathiTrust colleagues, Ryan Dubnicek and Boris Capitanu to produce a working dataset for topic modeling. *Fig.* 1 shows the steps to develop a working dataset for this project.

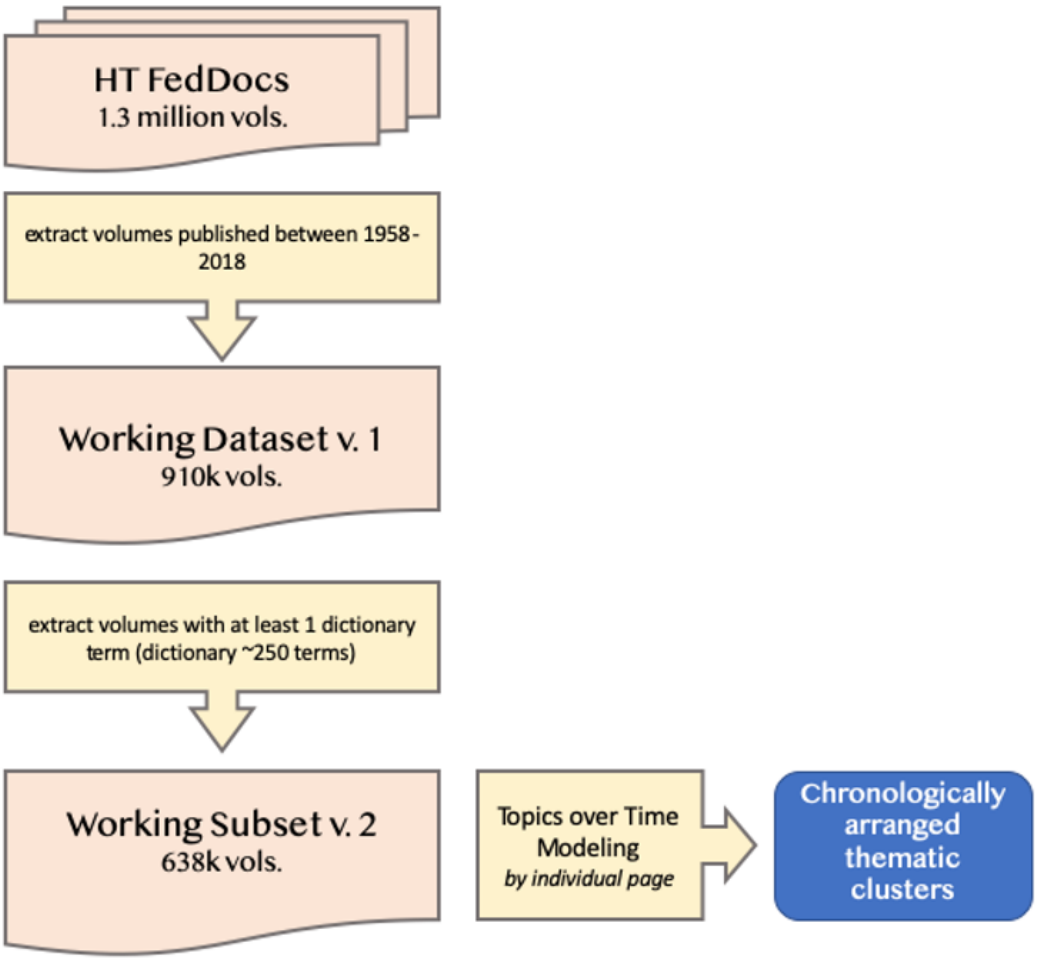

Figure 1. Creating our Working Dataset

The scoping of the working dataset first considered only those volumes identified within the U.S. Federal Documents Collection, which currently houses over 1.3 million volumes. Given this project's focus on federal policymaking in the latter half of the 20th century, we bounded our dataset by considering only those documents that were published between 1958 and 2018, reducing our dataset to ~910,000 volumes.

To further define our working dataset for our subsequent topic modeling, Castro Samayoa and Dubnicek received helpful guidance from Heather Christenson, program officer for federal documents and collections at HathiTrust, to understand some of the intricacies of working with the FedDocs collection. Christenson's guidance offered helpful insights on the limitations of using SuDocs due to their limited prevalence (~65% of the collection), and suggested OCLC numbers as a more helpful strategy, while also using the HathiTrust U.S. Documents Registry database to estimate the percentage of missing volumes from our collection. Given that FedDocs metadata is dependent on the quality of the input from submitting libraries, we learned to be cautious about the integrity of these data and have made a note to continuously assess the limitations of identifying edge cases due to metadata issues.

Alongside these insights from Christenson, Castro Samayoa developed a working dictionary of terms of interest relying on prior scholars' work detailing the history of ethnoracial terminology within the United States. Specifically, scholars of the U.S. census have chronicled the various shifts in terms employed to mark distinctions of race, ethnicity, and ethnonationalities within the United States (Anderson, 2016). This list was supplemented with additional terms used to describe a broader matrix of markers of social differences, including genders, sexes, sexualities, religions, and dis/abilities. The working dictionary yielded ~250 terms to use as a way of further culling the working dataset to those documents that explicitly named social differences in its contents. Capitanu's expertise with regular expressions ensured that the terms listed by Castro Samayoa could be positively identified within our working dataset across varying spellings, tenses, and numbers.

By excluding any volume that did not have at least one of the terms from the dictionary, we reduced the working dataset to ~650k volumes. Our work ahead will examine the changes over time for these various terms and examine a more granular unit of analysis by delving into volumes' unique pages. Given the interest in developing trends over time with a specific focus when specific race-related terms were employed in federal documents, the use of Topics Over Time (TOT) is most suitable for this project (Wang & McCallum, 2006). Unlike other LDA approaches to topic modeling, TOT ensures that "parameter estimation is driven to discover topics that simultaneously capture word co-occurrences and locality of those patterns in time" (Wang & McCallum, 2006, p. 424).

Preliminary Contributions & Next Steps

The original scope of this project was targeted to federal documents focused on education. However, the support of HathiTrust colleagues has made it possible for the working dataset include all federal documents within the specified parameters (1958-2018). This will prove to be a valuable contribution to any scholar outside of education who is also interested in the way that U.S. federal documentation traces the shifts in ethnoracial rhetorics across multiple federal agencies.

References

Anderson, M. J. (2016). *The American Census: A Social History*. Yale University Press.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '06*, 424. https://doi.org/10.1145/1150402.1150450