Check for updates

SOCIUS
Data Visualization

Socius: Sociological Research for a Dynamic World
Volume 4: 1–2
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2378023118809264
srd.sagepub.com
SAGE

Text Analysis with JSTOR Archives

John A. Bernau[1]

Abstract

I provide a visual representation of keyword trends and authorship for two flagship sociology journals using data from JSTOR's Data for Research repository. While text data have accompanied the digital spread of information, it remains inaccessible to researchers unfamiliar with the required preprocessing. The visualization and accompanying code encourage widespread use of this source of data in the social sciences.

Keywords

visualization, text analysis, sociology

The modern proliferation of digital data is well documented (Bail 2014). Large repositories of digital text have similarly revolutionized the possibilities for quantitative research in the social sciences (Evans and Aceves 2016). However, text data often require extensive processing and manipulation to get into a usable format, deterring otherwise inspired researchers. Using modern repositories like JSTOR's Data for Research (https://www.jstor.org/dfr/), one can easily download preprocessed text data, opening up a wealth of possibilities for cutting-edge scholarship.

For example, any user on JSTOR can create and request a data set that includes meta-data and word frequencies for up to 25,000 articles at a time. Data sets are based on search parameters that allow filtering by keyword, publication type, or specific journal titles. Figure 1 presents two analyses using this repository. The left pane of Figure 1 uses every research article published in the *American Sociological Review* between 1936 and 2015 (N = 5,320) to plot the use of three keywords over time: *race, class*, and *gender*. Yearly totals are calculated from individual article word frequencies, plotted on a log2 scale, and fit with a smoothed trend line. This is a simple and effective way to observe longitudinal trends in word usage. Among the "holy trinity" of the social sciences, *class* appears to be a mainstay of sociological conversations, whereas *race* and *gender* gained increased attention post-1960.

The right pane of Figure 1 depicts every research article published in the *American Journal of Sociology* between 1897 and 2014, plotted according to page length and publication date (N = 5,060). Trend lines are based on linear estimates of page length as a function of date and number of authors. This visualization presents a clear parabolic trend in article length and the significant effect of co-authorship over time (R^2 = .35). Individual article points are sized and colored according to residuals, emphasizing outliers. Albion Small's 150-page editorial retrospective on the discipline in 1916 (an extreme outlier) is given a text label. This is a simple and effective way to observe longitudinal trends in discipline norms surrounding article length and collaboration.

Equipped with similar data, researchers can explore countless other substantive questions. Using authorship data, one could examine the diffusion of collaborative publishing models across different disciplines. Using text data in JSTOR's bag-of-words format, one could easily conduct dictionary-based text analysis on key terms or more sophisticated topic modeling to examine the themes that characterize sociological research and their fluctuations over the past hundred years. A similar research agenda could provide insight into the degree to which themes in sociology overlap with related disciplines like political science, anthropology, or psychology. Paired with author affiliations, one could even examine the institutional or regional concentration of substantive research areas over time, giving a firm empirical foundation to future discussions of theoretical "schools" of thought. For an example of recent bibliometric research, see Borrett et al. (2018).

The plot was produced using R and ggplot2 (R Core Team 2018; Wickham 2016). Code used to produce the plots,

[1]Emory University, Atlanta, GA, USA

Corresponding Author:
John A. Bernau, Department of Sociology, Emory University, 1555 Dickey Dr., 225 Tarbutton Hall, Atlanta, GA 30322, USA.
Email: john.bernau@emory.edu

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

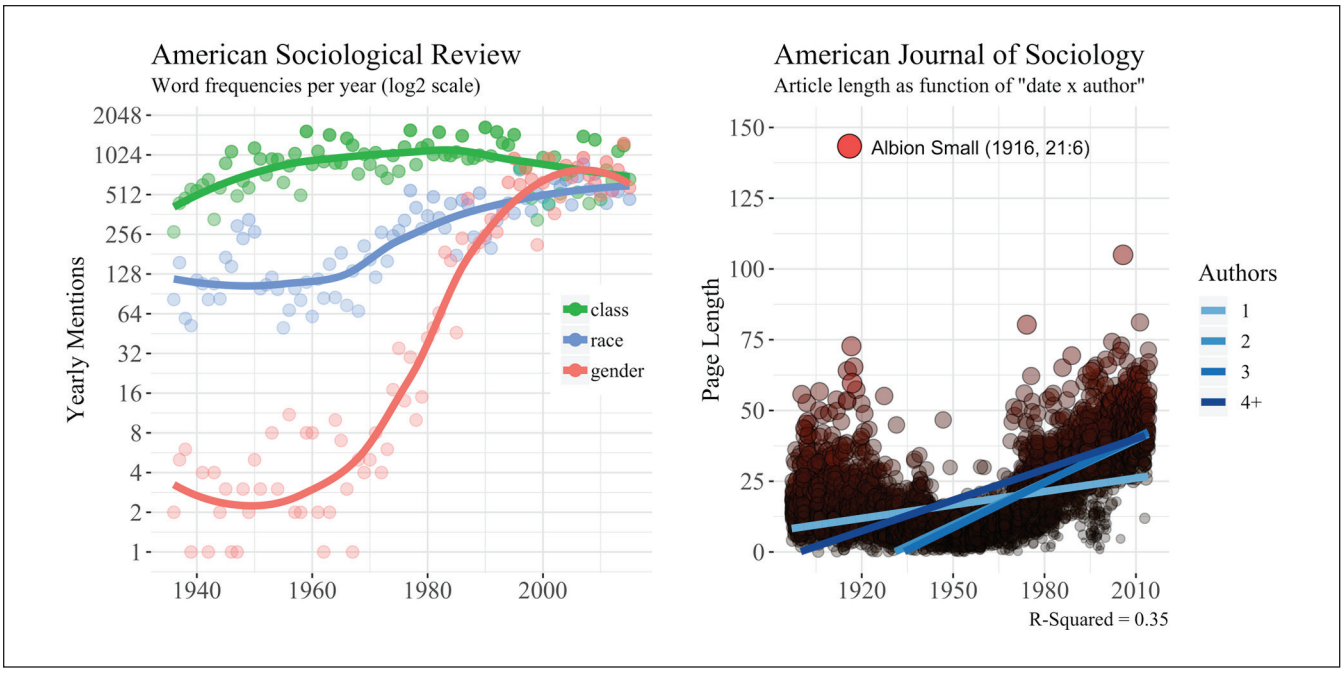

Figure 1. Trends in American sociological publishing. The left pane depicts annual word frequency totals for three key terms in the *American Sociological Review*'s publishing history, with yearly totals represented as points and summarized with a smoothed trend line. The right pane depicts article length for *American Journal of Sociology* articles between 1897 and 2014. Trend lines are based on predicted values of article length as a function of date, number of authors, and their interaction (date × number of authors). Points represent individual articles and are sized/colored according to residuals to emphasize outliers.

including instructions on converting JSTOR's xml metadata to a useable format, are available at https://github.com/johnbernau/jstor_dfr.

Supplementary Material

Supplementary material is available for this article online.

References

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43(3–4):465–82.

Borrett, Stuart R., Laura Sheble, James Moody, and Evan C. Anway. 2018. "Bibliometric Review of Ecological Network Analysis: 2010–2016." *Ecological Modelling* 382:63–82.

Evans, James A., and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42:21–50.

R Core Team. 2018. *R: A Language and Environment for statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Author Biography

John A. Bernau is a PhD candidate at Emory University. His research examines how groups use language to solve social problems. His dissertation uses recent methods in computational social science to examine American discussions of death and dying over the past 50 years.