Distant Horizons: Digital Evidence and Literary Change

By Ted Underwood. Chicago: University of Chicago Press, 2019.
200 pp. Softcover and EPUB. $27.50. Softcover ISBN 978-0-226-61283-6;
EPUB ISBN 978-0-226-61297-3.

Literary scholars' mining of libraries and archives will never cease, but the methods they use for their explorations of archival holdings are changing. The application of technological tools and algorithmic methods to humanities inquiry has risen in prominence over the past decade and loosely coalesces in a field frequently termed "digital humanities" or, more broadly, "digital scholarship." Now, what was once a specialized method of scholarship on the fringes of humanities research is an interdisciplinary catalyst driving new programs and areas of study in the humanities. Distant reading, also known as text data mining, is among the approaches to computational analysis in humanities that rely on digital collections derived from digitized holdings of academic libraries and archives. In this vein of research, Ted Underwood leads his readers on an introspective journey into the new horizons of literary analysis that can be unearthed through text mining. Underwood, professor of English and information sciences at the University of Illinois at Urbana-Champaign, is a leading scholar in applications of text data mining tools and methods for literary analysis, and he has published extensively in numerous articles and a previous book, *Why Literary Periods Mattered* (Stanford University Press, 2016). In his newest book-length work, *Distant Horizons: Digital Evidence and Literary Change*, Underwood shares his findings from the application of quantitative methods to texts and, in doing so, provides cogent insights into how numbers and prose are not in opposition after all.

Underwood circumvents the digital versus analog dichotomy by instead proposing a methodology he calls the "perspectival model." He explicates this methodology of statistical modeling by exploring three selected areas of investigation for literary analysis: analysis by genre, analysis of prestige, and analysis of gender representation. Throughout his analyses and posited findings, he consistently advances the complexly layered framework of statistical models that encompasses the seemingly objective quantitative measures and subjective interpretative facets. As he posits in the opening preface, "Quantitative models are no more objective than any other historical interpretation; they are just another way to grapple with the mystery of the human past, which doesn't become less complex or less perplexing as we back up to take a wider view" (p. xix). The subsequent analyses are consistent with this highly nuanced perspective on distant reading and text analysis.

When examining genres, Underwood begins with library classifications, most prominently through Library of Congress call numbers, and trains his computational models within the method of machine learning—to initially define genre by the patterns seen among a selection of texts. But when classifying additional texts, we learn that human interpretation of genre and the computational interpretation of Underwood's findings interestingly highlight that the shared patterns among the texts for a genre are not just surface markers of word type, but also are more subtly buried in themes.

His analysis of prestige takes a multifaceted approach that subtly pivots from analyzing the prestige of literary works across the centuries to actually revealing the complex nature of how both scholarly critics' and the public's views of prestige gradually but definitively shift. His analysis is particularly notable at revealing the types of questions that text analysis actually unearths: it is not really a matter of numbers to precisely chart what is a prestigious novel or readership, but rather, interpreting the patterns in tone and word choices.

Gender representation in Underwood's text analysis similarly pushes key questions to the forefront of how gender is even defined. The data visualizations in this chapter are particularly striking in showing the fluctuation of women authors and how women are represented in novels. And, overall, the two-tone data visualizations are effectively integrated with the text, so that I read, returned to browse the visualizations, and reread, making for a reading experience unique in its effective interchange between the computed visualizations and the text.

An awareness of his audience threads throughout Underwood's writing. The targets of his writing appear to be skeptics of the value of applying quantitative methods to literary analysis and also exploratory scholars who are intrigued but still at a loss as to how quantitative methods and texts can work together. Throughout, he firmly frames his distant reading analysis with critical theory that provides insights not only for literary scholars, but also those of us seeking to learn how text mining informs scholarly research practices. For archivists, of particular interest would be his use of digital libraries to build the body of works he analyzes, called a *corpus* in text-mining research.

Underwood worked intensively with the HathiTrust Digital Library (HTDL) to build the corpus of tens of thousands of texts for his research,[1] drawing upon the HathiTrust's digital collection of over 17 million volumes gathered from major academic libraries primarily in the United States and Canada, with a handful of international libraries as well. (As a matter of disclosure, this reviewer worked with Underwood on some of his early work with the HTDL.) The scope and limitations of the HathiTrust corpus are delineated in Underwood's analysis, but it is still worth noting that the HathiTrust digital collections originated from the volume of books, journals, and other materials that major research libraries deemed worthy of collecting over approximately the past century. While this is indeed a vast array of content, we must always consider the limitations: what is available in digital libraries today compared to what was actually the full scope of published works?

As collectors and curators of a number of the works eventually digitized for dissemination through digital libraries and archives, archivists are well advised to heed how these digital collections are being used, not only for reading access to collection materials, but for new, complex avenues of scholarly analyses. What can archives do to attend to the types of materials we make available as data sets?

Recent research, such as Thomas Padilla's *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*, Greta Bahnemann et al.'s *Transforming Metadata into Linked Data to Improve Collections Discoverability*, and Ryan Cordell's *Machine Learning + Libraries: A Report on the State of the Field* for the Library of Congress, provide synthesized theoretical and practical frameworks for archives and libraries to both contextualize the digitization work they have long done and forge new pathways to leverage their digitized holdings into machine-actionable data for research.[2] Furthermore, the growing body of scholarship investigating and exposing the need for decolonization of archives, critical librarianship, and nonneutrality in libraries and archives compels all of us to critically consider what gaps exist as our digital libraries and archives are increasingly used for new fields of literary analysis.[3] This issue is made visible in practical ways in this book, as Underwood's in-depth explanations of his data sets and his methods, both throughout the book and in a final appendix, are balanced with a full accounting of what is included and notable exclusions in the millions of texts he was able to obtain and analyze.

I appreciate how Underwood acknowledges the limitations of the corpora he could access, though his corpus is far more comprehensive than many other text-mining research projects due to his multiple collaborations. For future scholarly research, progress is being made to expand the digital libraries and archives available to all. The idea of "collections as data" has taken hold, and increasing numbers of archives and libraries are exploring how to increase the computational accessibility of digital collections. The Europeana Newspapers project; the HathiTrust Research Center; Always Already Computational: Collections as Data and Collections As Data—Part to Whole initiatives; the Library of Congress's Computing in the Cultural Heritage Cloud project; and numerous other large and small projects by archives, museums, and libraries together are building a formidable set of strategies and methods that cannot be ignored for how we can expand the breadth and scale of digital collections generated from our archival holdings to be accessible for computational analysis.[4] We also can explore the exponentially growing research outputs across the disciplines in the social sciences as well as the humanities that are drawing upon digital archives.[5]

Distant Horizons offers a meaningful and approachable study for understanding the scholarly impact and application of digital libraries and archives: its deceptively short 200 pages are dense with a perceptive and thoughtful literary examination combined with a notably approachable explanatory framework for text analysis and methods such as machine learning. While the level of the text makes it beneficial for the reader to have at least a passing familiarity with text analysis and text data mining, *Distant Horizons* overall speaks to a multiplicity of readers in staking out text analysis and digital scholarship as neither side avenues nor passing trends, but as an established path that integrates not just digital tools but also cross-disciplinary methodologies like statistical models to deepen the questions we can ask in humanistic inquiry.

© **Harriett Green**
Washington University in St. Louis

Notes

[1] "HathiTrust Digital Library," https://www.hathitrust.org. captured at https://perma.cc/R9TH-9SYK.

[2] Thomas Padilla, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries* (Dublin, OH: OCLC Research, 2019), https://doi.org/10.25333/xk7z-9g97; Greta Bahnemann, Michael Carroll, Paul Clough, Mario Einaudi, Chatham Ewing, Jeff Mixter, Jason Roy, Holly Tomren, Bruce Washburn, and Elliot Williams, *Transforming Metadata into Linked Data to Improve Digital Collections Discoverability: A CONTENTdm Pilot Project* (Dublin, OH: OCLC Research, 2021), https://doi.org/10.25333/fzcv-0851; Ryan Cordell, *Machine Learning + Libraries: A Report on the State of the Field* (Washington, DC: Library of Congress, 2020), https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf, captured at https://perma.cc/YW3Q-JWM6.

[3] Ricardo Punzalan and Michelle Caswell, "Critical Directions for Archival Approaches to Social Justice," *Library Quarterly* 86, no. 1 (2016): 25–42, https://doi.org/10.1086/684145; nina de jesus, "Locating the Library in Institutional Oppression," *In the Library with the Lead Pipe,* September 24, 2014, http://www.inthelibrarywiththeleadpipe.org/2014/locating-the-library-in-institutional-oppression, captured at https://perma.cc/Z9ER-27Y3; Chris Bourg, "Never Neutral: Libraries, Technology and Inclusion," *Feral Librarian* (blog), January 28, 2015, https://chrisbourg.wordpress.com/2015/01/28/never-neutral-libraries-technology-and-inclusion, captured at https://perma.cc/9ZRT-EFU7.

[4] Nataa Daki and Aleksandra Trtovac, "Historical Newspapers Content as a Base for Scientific Research—Europeana Newspapers Project," *Proceedings of IFLA 2014, 13–14 August 2014, Geneva, Switzerland* (IFLA, 2014); Thomas Padilla, Hannah Scates Kettler, Stewart Varner, and Yasmeen Shorish, "Collections as Data: Part to Whole," OSF Home, December 4, 2020, https://osf.io/r9n3s; Thomas Padilla, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner, *Final Report: Always Already Computational: Collections as Data*, Zenodo, May 22, 2019, https://doi.org/10.5281/zenodo.3152935; HathiTrust, "Our Research Center," https://www.hathitrust.org/htrc; Library of Congress Labs, "Computing Cultural Heritage in the Cloud," https://labs.loc.gov/work/experiments/cchc, captured at https://perma.cc/WL5Z-WAEB.

[5] Nicole M. Brown, Ruby Mendenhall, Michael Black, Mark Van Moer, Karen Flynn, Malaika McKee, Assata Zerai, Ismini Lourentzou, and Cheng Xiang Zhai, "In Search of Zora/When Metadata Isn't Enough: Rescuing the Experiences of Black Women through Statistical Modeling," *Journal of Library Metadata* 19, nos. 3–4, 2019, https://doi.org/10.1080/19386389.2019.1652967; *Journal of Cultural Analytics,* https://culturalanalytics.org, captured at https://perma.cc/J5F3-UC3V.