Critical Practice in Text Data Mining Research Cluster, 2020-2021
Project Report

Submitted to the Humanities Research Institute June 7, 2021 by:

Spencer D. C. Keralis, Ph.D., Assistant Professor and Digital Humanities Librarian, University Library, spencerk@illinois.edu

Project Abstract
Text data mining (TDM) is the computational and statistical analysis of large corpora of texts. Often positioned in a dialectical relationship with, if not in opposition to, the humanities methods of aesthetics and hermeneutics, text data mining is one of the cornerstone methodologies of the broad and idiosyncratic field known as digital humanities. As a field that is necessarily allied closely with the computer and information sciences, the digital humanities, and text mining in particular, are implicated if not complicit in the problems of bias; representation in terms of gender, race, and class; labor ethics; and other problems that are endemic in the tech industries. This research cluster proposes to examine how text data mining, as a disciplinarily diverse field, has manifested these problems, and how the University of Illinois digital scholarship community can work together to address them, moving toward a critical practice of text data mining that is ethical, just, and inclusive.

Project Team
Project Lead & Principal Investigator
Spencer D. C. Keralis, Scholarly Communication and Publishing, University Library

Speakers
Nicole Brown, Saint Mary's College of California
Heather Froehlich, Penn State University
Jo Guldi, Southern Methodist University
Amanda Henrichs, Amherst College
Spencer D. C. Keralis, University Library
Mary Borgo Ton, Scholarly Communication and Publishing, University Library
Glen Worthey, HathiTrust Research Center, School of Information Sciences

Other Participants
Ryan Dubnicek, HathiTrust Research Center, School of Information Sciences
John Gallagher, English
Merinda Kaye Hensley, Scholarly Commons, University Library
Carissa Phillips, Research and Information Services, University Library
Dan A. Shalmon, Cline Center for Advanced Social Research
Dena Strong, Research Information Technology
Dan Tracy, Scholarly Communication and Publishing, University Library
Ted Underwood, School of Information Sciences, English

COVID-19 Impact Statement
This project was undertaken during the COVID-19 pandemic of 2020-2021. All participants were working remotely, and all meetings and events were conducted via Zoom. In this time of collective trauma and
heightened anxiety, the project added additional stress and pressure for all participants, particularly the project lead and presenters. As a result, we did not meet collectively as frequently as we might have under pre-pandemic conditions and chose to limit the scope of our inquiry – focusing on three topic areas rather than the six described in the proposal and curtailing some activities. Due to Zoom fatigue and other factors, attendance at some events was lighter than we had hoped. That said, at each event there were participants from beyond the UIUC campus community, which would not have been possible with in person events, and at every event we were gratified by the friendliness and robustness of the discussion.

Summary of Activities
In Fall semester of 2020, the research cluster conducted three workshops and one presentation. Lauren Klein and Catherine D’Ignazio, our original planned speakers, were invited to present for UIUC as part of the Just Infrastructures Speaker Series, so we pivoted in Spring semester 2021 to invite four speakers to present virtually, in lieu of a single capstone event.

Marketing
All events were publicized on the Campus Humanities Calendar, via the Digital Humanities listserv, and the Library News internal listserv with a request that Subject and Liaison Librarians share announcements with relevant departments and individual faculty. Workshops were cross promoted on the Savvy Researcher webpage and calendar.

Events
Recordings of select events are shared with the presenters’ permission on the CTDM Channel on Illinois MediaSpace. To allow participants to learn comfortably, workshops were not recorded.

An Introduction to Text Data Mining: Principles and Practices with Dr. Spencer D. C. Keralis. September 15, 2020 (Savvy Researcher Workshop)
Text data mining (TDM) is the computational and statistical analysis of large corpora of texts. In this entry-level workshop, participants will get an introduction to the general principles of TDM and learn about a few of the many tools and methods for quantitative analysis of text data, including word frequency, topic modeling, and sentiment analysis.

Representation in the Digital Record with Glen Worthey. October 30, 2020 (Savvy Researcher Workshop)
How diverse, inclusive, and representative are our digital library collections — and our library collections in general — and how can we know that? How and why does this matter? What can we do about it, and what are we doing about it?

Where to find millions of books, and how to read them with Glen Worthey. November 4, 2020
The HathiTrust Digital Library is the world’s largest, holding nearly 17.5 million digitized volumes from research libraries around the world. The HathiTrust Research Center, co-located here at UIUC and at Indiana University, is the research gateway to that enormous, near-universal library. This talk briefly introduced the Center’s text and data mining tools, which allow researchers to computationally analyze the millions of volumes of text held in the HathiTrust Digital Library.

Introduction to Topic Modeling with Dr. Mary Borgo Ton. November 12, 2020 (Savvy Researcher Workshop)
Do you have too many texts and not enough time to read them closely? Curious to see how a theme or an idea changes over time, by author, or by genre? Looking for patterns and themes but not sure where to start? This workshop was an introduction to topic modeling, an algorithm-based method for identifying clusters of words that appear together in a corpus of text. As we explored sample data sets, including the letters of Alexander Hamilton, we discussed ways to tailor this method to individual research interests and teaching.

**The Human Dimensions of Reading at Scale** with Dr. Amanda Henrichs and Dr. Heather Froehlich. March 19, 2021
In this talk, Drs. Henrichs and Froehlich reflected on the impact of the Postdoctoral Laborers Bill of Rights. They discussed postdocs as a vocational practice, and how they have applied (or not) the Bill of Rights in their own roles, and what that means in practice.

**My Business, My Burden: Reflections on ‘Critical Computation’ and other epistemological endeavoring through methods** with Dr. Nicole M. Brown. April 30, 2021
During this talk Dr. Brown offered some reflections on how she wrestles with the tensions embedded in the quantitative/qualitative dichotomy that persists within the discipline of sociology. She shared how her work to undo the erasure of Black women's contributions within the (archival and academic) digitized record offers one demonstration of the power of Black feminist interventions. Her (in progress) manuscript, *We Are Each Other’s Business: Black Women’s Intersectional Consumerism during the Chicago Welfare Rights Movement*, is her attempt to center the experiences of Black women activists while also challenging epistemological claims that are perpetuated through methods.

**The Dangerous Art of Text Mining** with Dr. Jo Guldi. May 21, 2021
In this talk, Dr. Guldi spoke from her work in progress, *The Dangerous Art of Text Mining*, in which she lays out objections to "prediction" in data science from historians, why history needs Natural Language Processing, and explores the hybrid world of temporal analytics emerging from interdisciplinary analysis, and more.

**Speaker Biographies**

**Dr. Nicole M. Brown** specializes in historical sociology, political consumerism, black feminisms and critical technology studies. Dr. Brown earned her PhD in Sociology from the University of Illinois at Urbana-Champaign (UIUC), and MSEd from Illinois State University. She has 15+ years of experience in student affairs and academic affairs. Her research and intellectual curiosities relate to consumerism, social movements, methods and methodologies (specifically archival and computational analyses), as well as knowledge production processes.

**Dr. Heather Froehlich** is the Literary Informatics Librarian at the rank of Assistant Professor at Penn State University (University Park, PA, USA). She was awarded her PhD and Masters of Research from the University of Strathclyde (Glasgow, UK), where she studied representations of social identity in Shakespeare and other Early Modern London plays; before that, she studied English and Linguistics at the University of New Hampshire (Durham, NH, USA). Her work draws heavily on corpus stylistics, historical sociolinguistics, literary linguistics, and digital humanities.

**Dr. Jo Guldi** is Professor of digital humanities and historian of political economy at Southern Methodist University. Dr. Guldi is co-author of *The History Manifesto*, which examines the impact of digital analysis...
on the discipline of history. Her work deploys sentiment analysis and topic modeling to examine the political and economic history of 18th century Britain and its empire. Dr. Guldi is also the designer and developer of *Paper Machines*, a free software toolkit for historians who wish to perform a “distant reading” of large-scale textual corpora.

**Dr. Amanda Henrichs** is a Visiting Assistant Professor at Amherst College, a small liberal arts college in Western Massachusetts. Since completing her PhD at Indiana University in 2016, she has held a series of postdocs and visiting appointments at IU and Amherst. Her work centers on the poetics of information management in seventeenth-century Renaissance humanism and twenty-first-century digital humanities. She is increasingly interested in data visualization as a tool for social change, including experimental visualization techniques that incorporate fine and fiber arts. She is a member of the Visionary Futures Collective, and a lead author on the Postdoctoral Laborers Bill of Rights (2019), which set out guidelines for employers and workers in the digital humanities.

**Dr. Spencer D. C. Keralis** is a scholar of the past, present, and future of the book. As the Founder and Executive Director of Digital Frontiers, Dr. Keralis promotes transmedia collaboration across disciplines by bringing together the makers and users of digital resources for humanities research, teaching, and scholarly communication. His current research examines labor ethics in digital humanities, and structural barriers to collaboration in the humanities. He currently serves as Assistant Professor and Digital Humanities Librarian with the University Library of the University of Illinois at Urbana-Champaign.

**Dr. Mary Borgo Ton** is the Digital Publishing Specialist at the University of Illinois Library. She received her Ph.D. in British Literature with a concentration in Victorian literature and a Graduate Certificate in Digital Arts and Humanities from Indiana University and has contributed to digital collections of materials from the global south, including Livingstone Online, One More Voice, and Archivo Mesoamericano. In her current position, she supports authors and editors in all stages of the publication process as they create long-form digital scholarly works in Pressbooks, Omeka, and Scalar for the Illinois Open Publishing Network.

**Glen Worthey** is Associate Director for Research Support Services in the HathiTrust Research Center (HTRC), based in the University of Illinois at Urbana-Champaign School of Information Sciences. Formerly, he was Digital Humanities Librarian at Stanford, 1997-2019, and was founding head of Stanford’s Center for Interdisciplinary Digital Research (CIDR). He’s held many roles in the international digital humanities community and is currently Chair-Elect of the Alliance of Digital Humanities Organizations (ADHO) Executive Board. His graduate work was in Russian children’s literature at the University of California, Berkeley.

**Summary of Expenditures**

All project funds were expended on speaker fees for guest speakers. Because no travel funds were necessary, we were able to virtually bring in 4 speakers and offer slightly higher honoraria than we would otherwise.

<table>
<thead>
<tr>
<th>Expense</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker Honoraria x4</td>
<td>$2500</td>
</tr>
<tr>
<td>Total Expenses</td>
<td>$2500</td>
</tr>
</tbody>
</table>
Key Observations

The Digital Record

The digital record is always incomplete. Gaps in the archival record create gaps in the digital record, which can distort results of text mining projects (as highlighted in talks by Worthey, Brown, and Guldi). Commercial vendors have limited data sets based on partner collections, or in the case of HathiTrust, collections that are drawn from multiple partners and for which the partners set digitization priorities based on what is available on the shelves at given moments. As Glen Worthey pointed out in his talk, this can result in gaps even in canonical works. For example, while HathiTrust holdings include some works by Octavia Butler, an important Black science fiction writer, key works by Butler are missing, likely because they were not available on the shelves when member institutions conducted digitization sweeps. This distorts not only the view of Butler’s work, but also the view of 20th century fiction overall, as it is available to researchers using HTRC tools. This is likely the tip of the iceberg of missing texts in HathiTrust collections, despite the suggestion of encyclopedism present in HathiTrust’s promotional language.

CTDM workshop presenters made clear efforts to model inclusive practice and to problematize the limits of digital archives. In her workshop on topic modeling, Mary Borgo Ton included a brief discussion of how corpus bias effects topic modeling results and gave examples of how to historicize and contextual topic models, specifically referencing Sara Duke’s project which used topic modeling to open queer readings of Alexander Hamilton’s letters to John Laurens. During the sandbox portion of the workshop, Ton shared sample corpora featuring material from the global south, including Amardeep Singh’s Colonial South Asian Literature and transcripts of videos from Archivo Mesoamericano, in an effort to push back against the dominance of Anglophone and Eurocentric corpora in DH. Ton’s workshop in particular promoted conversation about how to account for gaps in corpora, reckon with algorithmic bias, and advocate for ethical practices in DH, and clearly demonstrated how the workshop setting can become a locus for these discussions.

It behooves researchers to both recognize and acknowledge the gaps in the archives from which they are drawing, and to endeavor to understand how the limits of these archives limit in turn the evidentiary claims that can be drawn from text mining. As but one example, which came up in the Q&A for Dr. Guldi’s talk, primary source database vendors Gale and ProQuest have developed suites of text mining tools that draw from their primary source resources, which they offer to institutions (i.e. libraries) as high-dollar add-ons to database subscriptions; HathiTrust Research Center has a similar suite that draws from HathiTrust collections. On the surface, this seems great. Researchers can use the suites of tools – either open source tools like Mallet (Gale) or Jupyter notebooks with hooks into the primary source databases (ProQuest) – to conduct text mining research. But a scholar cannot search or analyze across company-owned data sets – a corpus can only be drawn from one company’s resources, and even then, only those resources to which an institution subscribes. As Alex Gil points out, this effectively means that “actual capital relations are driving the research questions,” with corporate vendors selling two interfaces to the same content. Cody Hennesy warns that this will “entrench unequal access for computational scholars across institutional lines,” as well-resourced institutions provide access to these add-on suites of tools, and less well-resourced institutions do not. This exacerbates the already problematic nature of the digital record, limiting research by vendor, by tool suite, and by institutional subscription or, in the case of HathiTrust, institutional membership. It is crucial that researchers work to
understand these limitations and clearly acknowledge the parameters of their data when publishing results.

**Algorithmic Bias**

As discussed by Brown and Guldi, and following the important work of Safiya Noble, Lauren Klein, and Catherine D’Ignazio among others, we must recognize that algorithms always carry the biases of their authors, in much the same way that other forms of written communication carry the biases of their authors. Code is not somehow intrinsic or pure, and the bias encoded in algorithms cannot be modeled away. The root cause, as described by D’Ignazio and Klein, is that information systems “are created by small groups of people and then scaled up to users around the globe. But those small groups of people are not at all representative of the globe as a whole, nor even of a single city in the United States” (28). The tech industries, and by extension the technical teams driving digital research in the humanities, tend to be dominated by affluent men, typically a “straight, white, cisgender man with formal technical credentials” – and data science tools built by these men will privilege the questions, perceptions and biases of these individuals. This is not of course to say that a queer person, a woman, or a Black, indigenous, or other person of color will not bring a particular set of perceptions or assumptions to code they write, but the absolute dominance of tech fields by a particular category of person has allowed the assumptions and biases of those people to propagate disproportionately within technical fields, with a global impact.

The questions we ask, the methods we use to ask them, the resources we use to answer those questions, the results we get from our queries, and the interpretation we make of those results are all shaped by gender, race, class, and other contingencies that we, frequently unconsciously, bring to our research. By carefully historicizing and contextualizing both our work and ourselves, by understanding the limits of our data including gaps that may be present, and by understanding the genealogy of any code we reuse or redeploy, we can to some degree mitigate bias, but we can never completely model it away. We must acknowledge this and center this knowledge in our interpretation of text data.

**Labor**

Labor in the academy is unequal, and digital projects can exacerbate and expose these inequities in ways that can be toxic for both research and pedagogy (Greenspan). This challenge has historically been addressed with “soft solutions” (Keralis, 2018) like the UCLA Student Collaborators’ Bill of Rights, the Postdoctoral Laborers Bill of Rights, and the Collaborators’ Bill of Rights, but these documents go only so far without buy-in from communities of practice or administration support in the form of policies and processes that endorse the values encoded in these documents.

The continuing reliance in the humanities on single-author credit for publications and funding impairs collaboration and can lead to the perception that humanists are undesirable or bad collaborators. As a result, students, technologists, and library faculty and staff are frequently leery of committing to labor on digital humanities projects without clearly defined roles and (Currier, Mirza, and Downing). Some of these anxieties can be addressed by formalizing relationships between scholars and DH laborers at all levels, including expectations for credit, authorship, and/or compensation, with Memoranda of Understanding, clearly documented processes and decision trees, and other agreements. These methods can slow the pace of interactions in order to allow all parties to think intentionally about their roles on a project and allow participants in DH research to center labor equity as new projects begin (Keralis, Mirza, and Seale).
Conclusion
Whenever we put the label “critical” on a thing in the academy we risk creating an environment where we are preaching to the converted. Those who are inclined toward or benefit from inequitable labor practices, who are incurious or disbelieving about algorithmic bias, or whose work focuses exclusively on nationalistic and canonistic collections well represented in the digital record, are unlikely to step into spaces in which those practices and assumptions are exposed to criticism. As such I do not feel that we were successful in our aim to develop a community of practice focused on critical text data mining, except insomuch as the collaborative relationships between some of the core project team were strengthened. That said, the conversations fostered by this research cluster established a baseline expectation for future conversations in the emerging digital humanities community on campus that we will focus on questions of equity, justice, and inclusion, not just as part of our research, but as part of our praxis. The library and librarians alone cannot drive or compel these conversations and cannot dictate research or pedagogic methods or professional values to research and teaching faculty. Conversely, research and teaching faculty must recognize the needs of technology and research support staff, and library faculty and staff for credit and compensation that conforms to the unique requirements of promotion, tenure, and retention standards for these job classes, and the values that inform their work. Any future community of practice – whether driven by a research cluster or other activity – must necessarily include representatives of research and teaching faculty, technology and research support staff, and library faculty and staff as co-equal partners.

We are grateful to the Humanities Research Institute for their support of these discussions in these challenging times.

Bibliography


Gil, Alex. “That us too at @columbialib. Check how complicated this is. We just got a new ProQuest product for doing Text Data Mining. ...” [5 Tweet thread] 3:27 PM · May 18, 2021
https://twitter.com/elotroalex/status/1394751394767380481?s=20

https://dhdebates.gc.cuny.edu/

Hennesy, Cody. “It boggles my mind that vendors get away with charging libraries for the *exact* same content twice: once for search/retrieval, and again for TDM. It will entrench unequal access for computational scholars across institutional lines.” [Tweet] 10:41 AM · May 19, 2021
https://twitter.com/codyhennesy/status/1395041835395862530?s=20

