IFLA International News Media Conference

"Collecting, Preserving, and Transforming the News – for Research and the Public"
27-28 April 2017 – Landsbókasafn Íslands-Háskólabókasafn (The National and University Library of Iceland), Reykjavik, Iceland

What Can Text Mining Reveal about the Use of Newspapers in Research?

Mary Feeney
Research & Learning Department, The University of Arizona Libraries, Tucson, AZ, USA
mfeeney@email.arizona.edu

Abstract:

Newspapers are used in many ways by researchers. By examining scholarly literature, one can learn more about which newspapers are used and the disciplines of scholars who use them. This examination can be facilitated by JSTOR Data for Research (DfR), a free tool that enables text mining of the journal articles in the JSTOR archive. Using JSTOR DfR, this paper studies the occurrence of specific newspaper titles – from large national newspapers to local publications – in scholarly literature. The subjects and disciplines, key terms, and word frequencies associated with the journal articles that use these newspapers are also explored.

Keywords: Text mining, JSTOR Data for Research, Newspapers, Researchers

Introduction

In "The Many Uses of Newspapers," Alison Jones provides an excellent summary of the varied ways that scholars in many disciplines have used newspapers in their research, from examining newspaper advertisements to study women's occupational opportunities, studying obituaries to understand social history, looking at political cartoons and their influence on public perceptions, and much more.[1] Historians, archaeologists, geographers, and genealogists were all found to have used newspapers in their research.[2] By examining scholarly literature, one can learn more about who uses newspapers and what topics they write about when using them. Text mining is a method that can facilitate this examination, and JSTOR Data for Research (DfR) is one such tool that enables text mining of the journal articles in the JSTOR archive.[3]

For libraries, archives, and other cultural heritage memory institutions, a deeper insight into what scholars write about when using newspapers may help us focus on providing access to particular titles, promote alternative sources that have not been used as widely, determine

1 Alison Jones, "The Many Uses of Newspapers," *Technical Report for IMLS Project "The Richmond Daily Dispatch"* (2005). Retrieved from http://dlxs.richmond.edu/d/ddr/docs/papers/usesofnewspapers.pdf.
2 Ibid.
3 JSTOR Data for Research, http://dfr.jstor.org/.

which titles to digitize next, and provide tools that can help researchers better utilize the digital collections that are available.

Literature Review

Text mining is a "set of practices" that researchers may use "to approach text in new ways."[4] Corpora, such as books in the HathiTrust Library or digitized newspaper collections, allow texts "to become more like 'data;' that is, something that can be processed computationally."[5] One such corpus can be found in the JSTOR database, a multi-disciplinary digital archive of academic journals, books, and other materials. Started in 1995, JSTOR now has more than 10,000 institutions accessing the database, and the journal archives include over 2,000 titles and many millions of pages.[6] It is a widely-used resource of scholars in many disciplines.

In 2008, JSTOR launched its Data for Research (DfR) service, described as a "free data mining tool for journal content on JSTOR."[7] DfR allows scholars "to find useful patterns, associations and unforeseen relationships in the body of research."[8] It can be used by researchers interested in text mining the content in JSTOR, regardless of whether one's institution subscribes to the JSTOR database. A faceted search can be used to refine search queries, and data can be viewed across search results and at the individual document level. Users can create an account to download data sets, which include citations, word counts, n-grams, and key terms for up to 1,000 documents, for further analysis.

At the IFLA World Library and Information Congress in Lyon, Leonard described how libraries can support scholars in the digital humanities in mining digitized texts. Libraries can use existing tools, such as JSTOR DfR, or build tools to use with digital collections, and libraries should work to make sure we provide access to raw data, such as text files, to enable the kind of text and data mining that researchers want to do.[9] JSTOR DfR has "crucially…been at the forefront of thinking about how to make in-copyright material available" for text mining by making word counts "available on a per-article basis. This lets text-mining algorithms look for semantic patterns without giving away human-readable articles that would destroy the financial model behind JSTOR."[10]

Scholars in different disciplines have made use of this functionality of the DfR tool. Hundreds of data sets have been created and downloaded every year since DfR was introduced, and JSTOR has also worked with researchers to create and access larger data sets that cannot be downloaded directly from DfR.[11] There are several examples in the scholarly literature of research conducted with DfR. The review of the literature in this study focuses on *how* scholars used DfR and less on their results or the conclusions they drew.

---

[4] Devin Higgins, "Reading and Non-Reading: Text Mining in Critical Practice," in *Top Technologies Every Librarian Needs to Know: A LITA Guide*, ed. Kenneth J. Varnum (Chicago: American Library Association, 2014), 85.

[5] Ibid., 86.

[6] JSTOR, "New to JSTOR? Learn More About Us," accessed April 13, 2017, http://about.jstor.org/10things.

[7] Ibid.

[8] JSTOR, "About Data for Research (DfR)," accessed April 13, 2017, http://dfr.jstor.org/??view=text&&helpview=about_dfr.

[9] Peter Leonard, "Mining Large Datasets for the Humanities." Paper presented at: IFLA WLIC, 16-22 August 2014, Lyon, France. Retrieved from http://library.ifla.org/930/1/119-leonard-en.pdf.

[10] Ibid., p. 11.

[11] http://about.jstor.org/10things. Accessed April 13, 2017.

DfR enables researchers to examine changes in word usage over time. For example, “hath” declined in usage after 1900, and “chymistry” also dropped in usage until a later spike that may have been because of scholars citing the earlier works.[12] Other potential uses of DfR are for historians to trace “the spread…of theories…and ideas over time,” and as an advanced tool for scholars in bibliometrics.[13] After a detailed and useful explanation of how to use the features of the DfR interface, King et al discussed the results of a bibliometric study that examined the use of mathematical techniques in economics literature. The authors also explored the challenges in interpreting results from a DfR search and pointed out the problem of the rolling wall of coverage in JSTOR. Since JSTOR is primarily a journal archive, coverage of journals does not usually come up to the present, thus more recent scholarship will not be included, and the ending date varies by journal.

By searching the references in JSTOR publications, Hoyt showed that film studies scholars tend to use the same magazines in their research, specifically that four publications have been repeatedly cited more than others, while “half of all historic film periodicals have never been cited in JSTOR.”[14] One of the faceted search features is the ability to select publications in subject areas. Almost 14,000 articles classified as film studies journals were mined in DfR for this study.

DfR also allows researchers to select specific journal titles to search. Marshall searched for specific terms in DfR within two specific population studies journals, one British and one French, covering 1946-2005. The author’s interest was in looking at “cross-national differences in the development of academic disciplines.”[15] Marshall then exported the word counts as a data set and used them in topic modeling, “a method of textual analysis that searches for latent structures underlying sets of documents” and “allows the identification of topics in a set, or corpus, of documents.”[16]

Three folklore studies journals were selected in DfR as the focus of Laudun and Goodwin’s exploration to “map historical shifts” over 125 years in this discipline.[17] Almost 7,000 articles were mined to see what patterns would emerge from the texts and whether “the exploration turn up any under-recognized – even latent – dynamics or trends worth further consideration.”[18] Using data from DfR, the authors used topic modeling to draw out fifty distinct topics in the corpus. Similarly, Wang et al used DfR to text mine one specific journal for key phrases and then used topic modeling to identify over a dozen topics that “reveal the ebb and flow of consumer research topics” over forty years of the journal.[19]

---

[12] John Burns et al, "JSTOR-Data for Research," in *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Proceedings,* ed. Maristella Agosti et al (Berlin; Heidelberg: Springer, 2009), 419.

[13] Art King, Brian Simboli, and Kevin Rom, "JSTOR’s ‘Data for Research’: A Bibliometric Analysis of Mathematics in Economics," *Issues in Science and Technology Librarianship* (Fall 2012).

[14] Eric Hoyt, "Lenses for Lantern: Data Mining, Visualization, and Excavating Film History's Neglected Sources," *Film History: An International Journal* 26, no. 2 (2014), 148-151.

[15] Emily A. Marshall, "Defining Population Problems: Using Topic Models for Cross-National Comparison of Disciplinary Development," *Poetics* 41, no. 6 (2013), 702.

[16] Ibid., 706.

[17] John Laudun and Jonathan Goodwin, "Computing Folklore Studies: Mapping Over a Century of Scholarly Production Through Topics," *Journal of American Folklore* 126, no. 502 (2013), 457.

[18] Ibid., 459.

[19] Xin Shane Wang et al, “The *Journal of Consumer Research* at 40: A Historical Analysis,” *Journal of Consumer Research* 42, no. 1 (2015), 6.

Edelstein compared searches in DfR of the term "the Enlightenment" to national Enlightenment terms such as "Scottish Enlightenment," "German Enlightenment" (or *Aufklärung*), and "French Enlightenment" to see if the data could "shed any light on the practice of Enlightenment naming."[20] Data mining can "guide our analyses of secondary sources,"[21] and Edelstein noted this method is "one of few options for studying Enlightenment scholarship across multiple disciplines"[22]

Economics was another discipline explored using DfR. Kufenko and Geiger searched terms related to business cycles and economic crises in the DfR "Economics" and "Business and Economics" subjects, which include hundreds of thousands of articles. Along with other methods, they used the DfR data to examine "whether or not the economic literature on business cycles is correlated with movements and changes in actual economic activity."[23]

In combination with archival research, DfR was used to investigate how the field of demography was developed in conjunction with funding and institutional support, by looking at articles classified in DfR in the subject "population studies" in journals categorized in the discipline "social sciences." Merchant notes that articles in DfR are classified in two ways: "'subject' is an article-level classification determined by JSTOR through an LDA process; 'discipline' is a journal-level classification determined by the publisher of each journal."[24]

In the field of literary studies, Goldstone and Underwood used data from DfR for topic modeling "to elucidate historical patterns in a corpus of more than 21,000 scholarly articles from the last 120 years."[25] They chose seven literary studies journals with long runs for their DfR search and limited their results to focus on scholarly articles. The authors make a comparison to content analysis in the social sciences – the problem of "going from individual examples to interpreting on a larger scale"[26] – and emphasize the importance of contextualization when interpreting texts.

Scholars who have made use of DfR have been in a range of disciplines, from economics to literary studies, demography to film studies, and consumer research to folklore studies. They each approached the tool from different angles. Some selected to search a particular journal title or a few titles in a certain discipline, some searched key terms across an entire disciplinary group. Many wanted to examine the change in their discipline over time, and most used their resulting data from DfR in other ways, such as in topic modeling or in conjunction with other research materials.

These examples demonstrate the variety of possibilities in using the JSTOR DfR tool for text mining. Knowing that scholars in many disciplines use newspapers in their research, and given the broad range of disciplinary scholarly literature covered in the JSTOR archive, the

---

20 Dan Edelstein. "Enlightenment Scholarship by the Numbers: dfr.jstor.org, Dirty Quantification, and the Future of the Lit Review," *Republics of Letters: A Journal for the Study of Knowledge, Politics, and the Arts* 4, no. 1 (October, 2014), 15.

21 Ibid, 2.

22 Ibid., 5-6.

23 Vadim Kufenko and Niels Geiger, "Business Cycles in the Economy and in Economics: An Econometric Analysis," *Scientometrics* 107, no. 1 (2016): 43.

24 Emily Klancher Merchant, "A Digital History of Anglophone Demography and Global Population Control, 1915–1984," *Population and Development Review* 43, no. 1 (2017), 29.

25 Andrew Goldstone and Ted Underwood, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History* 45, no. 3 (2014), 360.

26 Ibid., 366.

JSTOR DfR tool was chosen for this study as a potential way to explore how scholars across disciplines use newspapers in their research.

Methodology

A list of newspaper titles with long runs was generated to include several national United States newspapers representing different regions of the country and two local newspapers in Tucson, Arizona, where the University of Arizona is located. Newspapers were excluded if their titles, including earlier titles, are very common and could refer to more than one newspaper, such as the *Daily Morning Chronicle* (a preceding title the *San Francisco Chronicle*) or the *Daily Constitution* (a preceding title for the *Atlanta Journal-Constitution*). Newspapers whose titles could be interpreted with other meanings within the texts in DfR, such as "the times" were not included to avoid false search results.

The titles used in this study were:

- *Arizona Daily Star*, published 1879-present
- *Boston Globe*, published 1872-present
- *Chicago Defender*, published 1905-present
- *Chicago Tribune*, published 1847-present
- *Los Angeles Times*, published 1881-present
- *New York Times*, published 1851-present
- *Tucson Citizen*, published 1870-2009 (print ceased)
- *Washington Post*, published 1877-present

The *U.S. Newspaper Directory, 1690-present* available from the Chronicling America site was consulted to verify title variations of the selected newspapers to ensure that searches would include all versions of the newspaper titles.[27]

Each newspaper title and its variants were searched in DfR. There is the ability to search "Anywhere in document," which is the default search, or to search in the title, author, abstract, caption, key terms, or references. Searching for scholarly articles with the newspaper titles in the abstracts only would have been too limiting. A search in the references was tested, but upon further investigation and comparing results to a full-text search, it was clear that newspapers are not always cited in references but rather in-text. Thus, searching "Anywhere in the document" was chosen for this exploratory research.

The faceted search was used to limit results to "Content Type – Journal" and "Article Type – Research article." The Subject Groups, Subjects, and Key terms associated with the search results for each newspaper title were downloaded for analysis. DfR also provides data for Year of Publication, but that data was not used for this analysis.

Results and Discussion

Not surprisingly, there were substantially more articles that included the *New York Times* than any of the other newspaper titles searched (Table 1):

[27] *U.S. Newspaper Directory, 1690-present*, accessed April 13, 2017, http://chroniclingamerica.loc.gov/search/titles/.

> …the *New York Times* is greatly overrepresented in its use as a historical source not only due to the fact that it has an extensive index but also due to the problem of presentism. Researchers project this paper's current importance into the past when there were actually other newspapers that served as the 'papers of record' for the time.[28]

Jones quotes from a journalism and media studies professor's presentation that "newspaper use in research is often based on availability, which privileges proximity and causes a 'certain imbalanced cosmopolitanism.'"[29]

Bingham, referring to *The Times* of London, also makes this point:

> …research may be distorted by the availability of certain titles and the absence of others. The attractiveness of working with digital archives means that many scholars will inevitably be drawn to those titles that they can access via their computer—even if they are not necessarily the most appropriate publications to use…In the long term, scholars must make the case for digitizing as many titles as possible, and encourage their libraries to subscribe to them…it is important that the convenience of digitization does not make us lazy researchers.[30]

While the number of articles referencing the *New York Times* is small relative to the full DfR corpus of research articles in journals (4,790,792), it is significantly more than the other newspaper titles searched.

Table 1. Number of articles using newspapers

New York Times	99,471
Chicago Tribune	27,023
Washington Post	25,477
Los Angeles Times	11,532
Boston Globe	5,600
Chicago Defender	1,039
Arizona Daily Star	585
Tucson Citizen	367

While this data shows the number of journal articles that include a newspaper title in comparison to other newspaper titles, it does not reveal to what extent a newspaper was used in a scholar's research. The numbers represent how many documents had the key terms in them, but as Edelstein points out, "the JSTOR results suggest that it is common for *many* scholars to include at least a *few* references to [the search terms] in their articles," but that consequently, "all documents are placed on the same footing: an article…that happens to contain a fleeting mention of [the search terms] counts equally" as a fuller, in-depth study.[31] Likewise, Kufenko and Geiger comment that "such a method of counting [word occurrences] does not differentiate—as a proper and informed reading of an article could—between how intensely a topic was discussed, with which intention this may have been done, whether a

[28] Jones, "The Many Uses of Newspapers," 36.

[29] Ibid.

[30] Adrian Bingham, "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians," *Twentieth Century British History* 21, no. 2 (2010), 229.

[31] Edelstein, "Enlightenment Scholarship by the Numbers," 4,7.

contribution was purely theoretical, etc."[32] Despite these limitations, the data from DfR may nonetheless provide some insight into newspaper use, in comparison to other newspapers.

Subject groups

For each set of search results, DfR provides a breakdown by "Subjects," "Subject Groups," "Disciplines," and "Discipline Groups." Disciplines and the broader discipline groups are "topics under which JSTOR covered journals have been categorized," while the subjects are "generated algorithmically."[33] The "Charts View" in DfR provides a graph of the subject group distribution of the search results. Articles can be assigned to more than one subject group, so totals in each group are more than the total number of articles.

For comparison, the distribution of subject groups in the corpus of "Research articles" (excluding the "Article types" of book reviews, miscellaneous, news, and editorials) in DfR is shown in Figure 1. Research articles that included the *New York Times* were predominantly in the broad subject group of Social Sciences, followed by Area Studies, then Humanities (Figure 2). A smaller number were, in rank order: Business and Economics, Science and Mathematics, History, Law, Medicine and Allied Health, Arts.

Figure 1. Distribution of subject groups of all research articles in JSTOR DfR

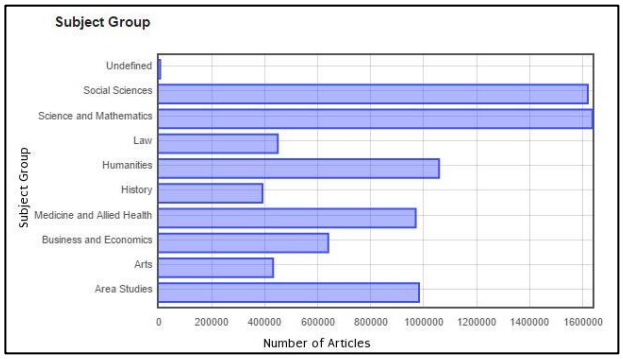
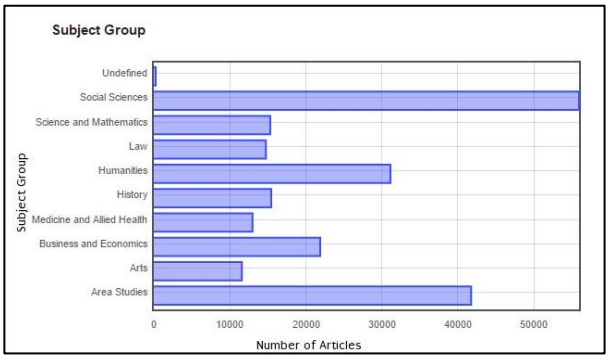


Figure 2. Subject groups of articles using the *New York Times*



Similarly, the largest subject group that included the *Chicago Tribune* was Area Studies, followed closely by the Social Sciences, then History and Humanities (Figure 3). Research articles that included the *Boston Globe* were also mostly in the Social Sciences, Humanities, and Area Studies (Figure 4). For articles using the *Los Angeles Times* (Figure 5), the top subject groups were Social Sciences, Area Studies, and Humanities, and for the *Washington Post* were largely in Area Studies and Social Sciences, followed by Humanities (Figure 6). Part of the reason for the similarities in subject groups for research articles using these five national newspapers is due to the newspapers occurring together in some articles. A search that includes all titles (and their variants) results in 185 articles.

[32] Kufenko and Geiger, "Business Cycles in the Economy and in Economics," 47.
[33] King et al, 3.

Figure 3. Subject groups of articles using *Chicago Tribune*

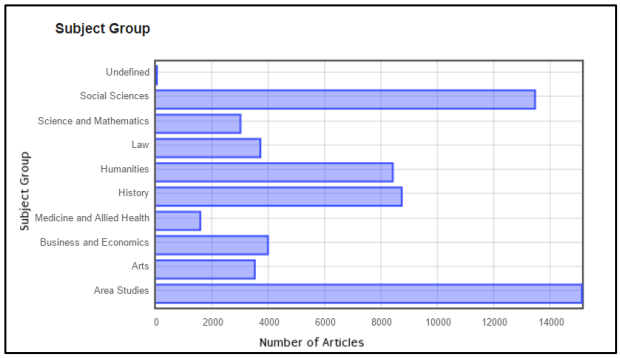


Figure 4. Subject groups of articles using *Boston Globe*

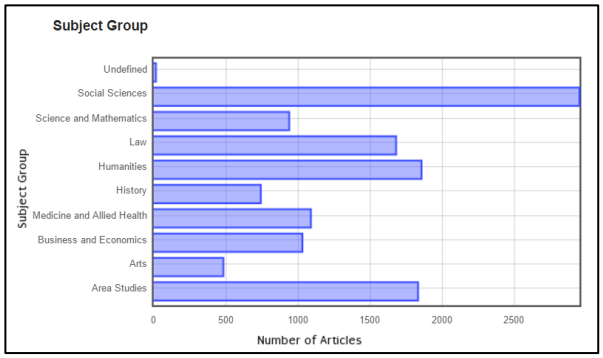


Figure 5. Subject groups of articles using *Los Angeles Times*

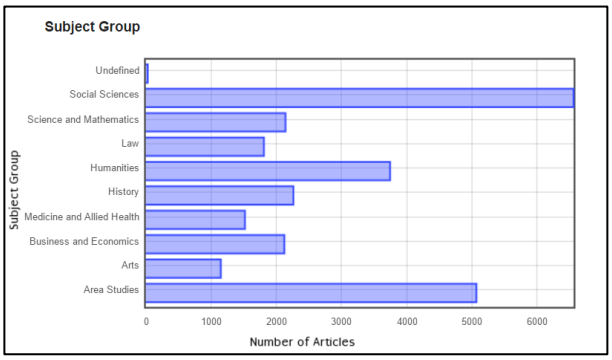
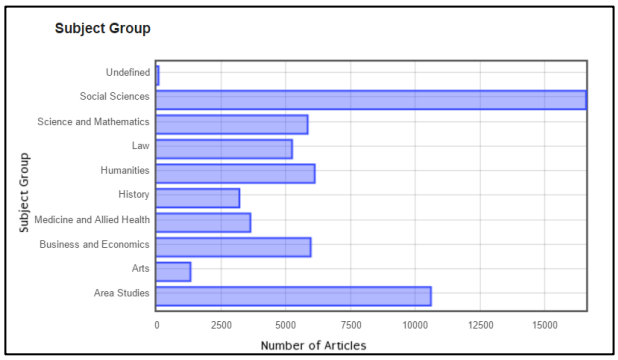


Figure 6. Subject groups of articles using *Washington Post*



In contrast, the highest subject groups for both the *Arizona Daily Star* and *Tucson Citizen* were History and Area Studies (Figures 7 and 8). The *Chicago Defender* had similar results, with Area Studies, History, and Humanities being the largest subject groups for the research articles using that newspaper (Figure 9).

Figure 7. Subject groups of articles using *Arizona Daily Star*

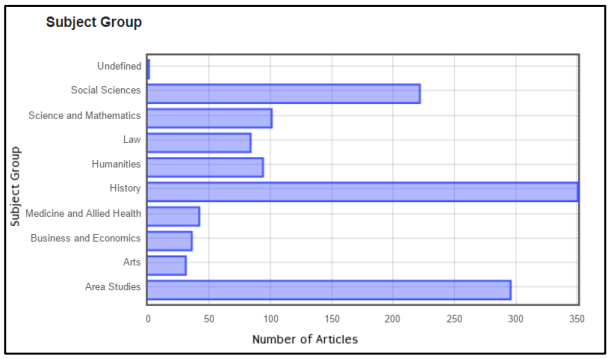


Figure 8. Subject groups of articles using *Tucson Citizen*

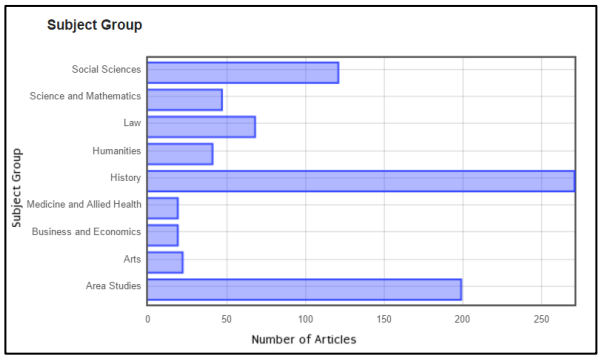
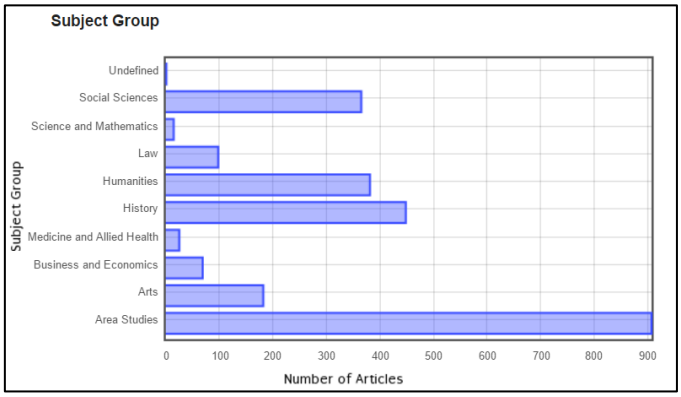


Figure 9. Subject groups of articles using *Chicago Defender*



Subjects

While the charts provided in DfR are for broad "subject groups," users can also download data for more granular subjects. Looking at the "subjects" data for the five national daily newspapers in this study shows Political Science as the highest or second highest subject for the search results for each of those newspapers. Law is second highest or highest for all but the *Chicago Tribune*, which had American Studies as its second highest subject (Table 2).

Table 2. Top Ten Subjects of Research Articles using National Daily Newspapers

New York Times	***Chicago Tribune***	***Washington Post***	***Boston Globe***	***Los Angeles Times***
Political Science	Political Science	Political Science	Law	Political Science
Law	American Studies	Law	Political Science	Law
American Studies	Law	Technology	Feminist & Women's Studies	Film Studies
Technology	African American Studies	Management & Organizational Behavior	Technology	Technology
Film Studies	Bibliography	Public Policy & Administration	American Studies	Feminist & Women's Studies
Social Work	Film Studies	Feminist & Women's Studies	Health Policy	American Studies
Management & Organizational Behavior	Slavic Studies	Middle East Studies	Social Work	Latin American Studies
Feminist & Women's Studies	Technology	Social Work	Public Policy & Administration	Social Work
African American Studies	Asian Studies	Health Policy	Management & Organizational Behavior	Management & Organizational Behavior
Performing Arts	Feminist & Women's Studies	African American Studies	Film Studies	Public Policy & Administration

Key terms

DfR also allows researchers to view "key terms" associated with their overall search results. Burns et al noted that "key terms for the entire data set" are "generated using TF/IDF"[34] King et al further explained that:

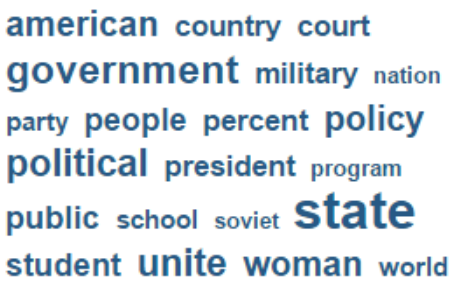
> …the frequency ranked 'Key Terms' are algorithmically weighted 1-grams: DfR sifts the 1-grams 'relevant' to a search, and presumably assigns the most relevant term a weight of one. Other terms are weighted relative to this term such that all have weight less than or equal to one. When one retrieves 'Key Terms' from the DfR interface, one receives this list of terms, absent the weights, aggregated over all publications returned by a search.[35]

Search results include a word cloud of the top key terms across the set of results (Figure 10). Each word cloud displays the top twenty-one key terms in each set of search results. The full list of key terms can be downloaded for each set of search results. For the large daily national newspapers – the *New York Times*, the *Washington Post*, the *Chicago Tribune*, the *Los Angeles Times*, and the *Boston Globe* – the top key terms are very similar. In fact, they share eleven of the top twenty key terms, not necessarily in the same rank order: state, court, woman, public, American, government, people, political, policy, unite, president. This may be due, in part, to the newspaper titles appearing together in some of the same research articles.

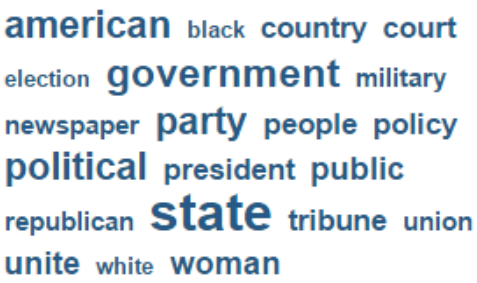
There are also some unique key terms in the top twenty. An obvious example is "Boston," which appears in the top ten key terms from research articles using the *Boston Globe.* But the only other occurrence of "Boston" in the full list of key terms for the other newspapers is in the *Chicago Tribune*, ranked about two hundred. "California," in the top twenty key terms for the *Los Angeles Times* search results, does not appear at all in the three hundred key terms for any of the other major newspaper title results. Other key terms that are in the top twenty for one newspaper and not others, but do occur somewhere in the three-hundred key term lists: newspaper (the *Chicago Tribune*), world (the *New York Times*), congress (the *Washington Post*).

Figure 10. Word Clouds Showing Key Terms in Articles Using National Daily Newspapers

New York Times

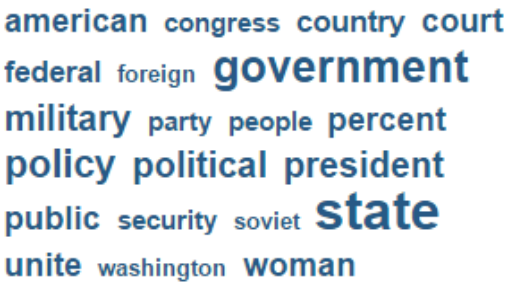


Chicago Tribune

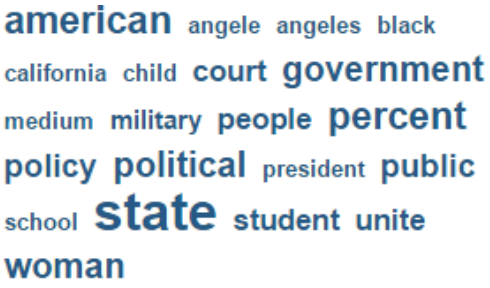


[34] Burns et al, 417.
[35] King et al.

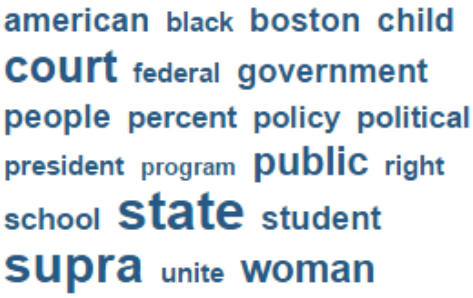
Washington Post



Los Angeles Times



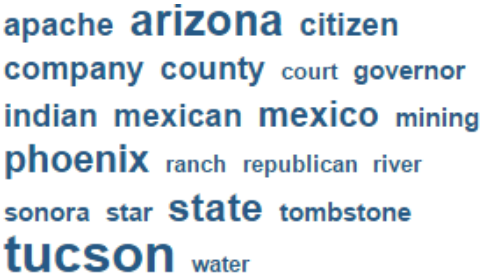
Boston Globe



While there are many similarities in key terms among the national newspapers, it is striking to see how the key terms compare to research articles that used local newspapers, the *Arizona Daily Star* and the *Tucson Citizen* (Figure 11).

Figure 11. Word Clouds Showing Key Terms in Articles Using Arizona Newspapers

Arizona Daily Star



Tucson Citizen

Several words appear only in the list of key terms for these two Arizona newspapers compared to the five national newspapers in this study, including Arizona, Tucson, Phoenix, Apache, Sonora, mining, and ranch. Several other terms in the Arizona newspapers' top twenty appear in the key terms for the other newspapers, but ranked much farther down the list: Mexico, Mexican, Indian, governor, border, water. Clearly, researchers who used these two local newspapers were much more focused on topics related to the region, that of our border with Mexico, American Indians, water issues, and mining and ranching that are major industries in the Arizona economy.

It is also apparent that the two word clouds for the Arizona newspapers are almost identical. A search in DfR with both *Arizona Daily Star* (and its variant titles) and *Tucson Citizen* (and

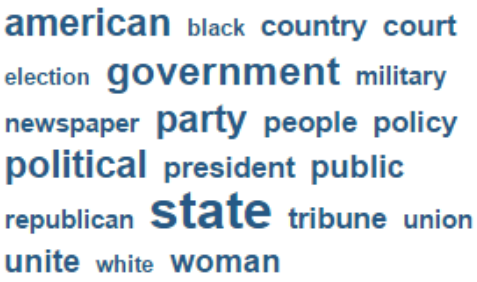
its variant titles) in the same search shows that the newspaper titles co-occur in about two hundred research articles. Researchers tended to use these two newspapers in tandem.

Another feature of JSTOR DfR is the ability to download a spreadsheet with the journal titles and number of articles in each journal from one's search. Viewing that list for the combined search of Arizona newspapers shows a narrow range of journals in which research articles using these two newspapers have been published. Almost eighty percent of the articles were published in only four journals: *Journal of Arizona History* (111 articles), *Arizona and the West* (30), *Journal of the Southwest* (14), and *Arizoniana* (10). The remaining articles were distributed in thirty-two other journal titles.

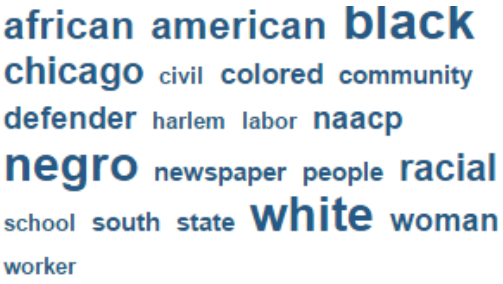
Another interesting comparison is that of the *Chicago Tribune* and the *Chicago Defender* (Figure 12). The *Chicago Defender* is a long-running weekly African American newspaper published in the same city as the major daily national newspaper, the *Chicago Tribune.* Many of the top key terms in research articles that used the *Chicago Defender* are not in the top key terms for the *Chicago Tribune* (NAACP, colored, Harlem) or are much farther down the list (African, racial, south, civil). On the surface, this may point to researchers who use the *Chicago Tribune* compared to the *Chicago Defender* writing about very different topics. The focus of the top key terms associated with the *Chicago Tribune* seem to be more political science in nature, but the broad subjects for both newspapers are generally similar (Figures 3 and 9).

Figure 12. Word Clouds Showing Key Terms in Articles Using Chicago Newspapers

Chicago Tribune



Chicago Defender



Conclusion

This exploratory study aims to investigate whether the text mining tool JSTOR Data for Research would be useful in understanding newspaper use in scholarly literature. The subject groups and subjects provide a broad view of what areas tend to use newspapers in their research. For the national newspapers used in this study, the Social Sciences, Area Studies, Humanities, and History have a higher occurrence of newspapers, but a whole range of other disciplines use newspapers, as well, from Business and Economics to Science and Mathematics, Arts, and Medicine. In contrast, the highest broad subject groups for two local Arizona newspapers and an African American newspaper in this study were History and Area Studies. The more specific subjects of research articles using the five national newspapers were most often Political Science, Law, American Studies, Technology, Feminist and Women's Studies, Film Studies, African American Studies, among others.

Given that there is some overlap in articles that included the individual national titles, a closer examination of the content of the articles would be useful. Hoyt, for one, advocates the combination of "close reading and archival research…with computational, 'distant reading' research methods."[36]

The key terms in articles using newspapers are only an initial and surface look at what scholars write about when newspapers are a part of their research. Future exploration of this data could include topic modeling of the key terms, as several other researchers have done with DfR data. It is also possible in DfR to download key terms and n-grams (bigrams, trigrams, quadgrams) at the article level, which would provide a more detailed look. There is an interesting contrast in key terms when comparing national newspapers to local newspapers. Understandably, the research articles that used the Arizona newspapers appear to focus on topics of local and regional importance in that state. An analysis of the search results in DfR of local newspapers from other parts of the country would be useful to see if this pattern holds. In addition, the brief comparison of the *Chicago Tribune* to the *Chicago Defender* points to researchers who use these two newspapers writing about very different topics. Further searching in DfR of other African American, ethnic, alternative, and local newspaper titles may help to expose whether researchers tend to utilize major national newspapers to the exclusion of a wider and more diverse range of newspapers.

In some ways, this exploration raises more questions than it answers, but one benefit of distant reading through text mining is perhaps to give direction to new areas of exploration and research. As Edelstein states:

> The numerical results are not meant to offer clear, definitive answers to particular questions, but simply, and in the best of cases, to provide a general, fuzzy sense of how, where, and when certain scholars have engaged with certain ideas. These are dirty data, and they're done dirt cheap: anyone, without any technical know-how, can download a mother lode of information. The trick, of course, is making sense of what it might mean.[37]

[36] Hoyt, "Lenses for Lantern," 155.
[37] Edelstein, 3.

References

Bingham, Adrian. “The Digitization of Newspaper Archives: Opportunities and Challenges for Historians,” *Twentieth Century British History* 21, no. 2 (2010): 225-231.

Burns, John, Alan Brenner, Keith Kiser, Michael Krot, Clare Llewellyn, and Ronald Snyder. "JSTOR-Data for Research." In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Proceedings,* edited by Maristella Agosti, José Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas, 416-419. Berlin; Heidelberg: Springer, 2009.

Edelstein, Dan. “Enlightenment Scholarship by the Numbers: dfr.jstor.org, Dirty Quantification, and the Future of the Lit Review.” *Republics of Letters: A Journal for the Study of Knowledge, Politics, and the Arts* 4, no. 1 (October, 2014): 1-26.

Goldstone, Andrew and Ted Underwood. “The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us.” *New Literary History* 45, no. 3 (2014): 359-384.

Higgins, Devin. “Reading and Non-Reading: Text Mining in Critical Practice,” in *Top Technologies Every Librarian Needs to Know: A LITA Guide*, ed. Kenneth J. Varnum (Chicago: American Library Association, 2014), 85.

Hoyt, Eric. "Lenses for Lantern: Data Mining, Visualization, and Excavating Film History's Neglected Sources." *Film History: An International Journal* 26, no. 2 (2014): 146-168.

Jones, Alison. “The Many Uses of Newspapers.” *Technical Report for IMLS Project “The Richmond Daily Dispatch.”* 2005. Retrieved from http://dlxs.richmond.edu/d/ddr/docs/papers/usesofnewspapers.pdf.

JSTOR. “About Data for Research (DfR).” Accessed April 13, 2017. http://dfr.jstor.org/??view=text&&helpview=about_dfr.

JSTOR. “New to JSTOR? Learn More About Us.” Accessed April 13, 2017. http://about.jstor.org/10things.

King, Art, Brian Simboli, and Kevin Rom. "JSTOR’s ‘Data for Research’: A Bibliometric Analysis of Mathematics in Economics." *Issues in Science and Technology Librarianship* (Fall 2012).

Kufenko, Vadim and Niels Geiger. “Business Cycles in the Economy and in Economics: An Econometric Analysis.” *Scientometrics* 107, no. 1 (2016): 43-69.

Laudun, John and Jonathan Goodwin. "Computing Folklore Studies: Mapping Over a Century of Scholarly Production Through Topics." *Journal of American Folklore* 126, no. 502 (2013): 455-475.

Leonard, Peter. “Mining Large Datasets for the Humanities." Paper presented at: IFLA WLIC, 16-22 August 2014, Lyon, France. Retrieved from http://library.ifla.org/930/1/119-leonard-en.pdf.

Marshall, Emily A. "Defining Population Problems: Using Topic Models for Cross-National Comparison of Disciplinary Development." *Poetics* 41, no. 6 (2013): 701-724.

Merchant, Emily Klancher. “A Digital History of Anglophone Demography and Global Population Control, 1915–1984.” *Population and Development Review* 43, no. 1 (2017): 83-117.

U.S. Newspaper Directory, 1690-present, accessed April 13, 2017. http://chroniclingamerica.loc.gov/search/titles/.

Wang, Xin Shane, Neil T. Bendle, Feng Mai, and June Cotte. “The *Journal of Consumer Research* at 40: A Historical Analysis.” *Journal of Consumer Research* 42, no. 1 (2015): 5-18.