Submitted on: 6/23/2014

Mining large datasets for the humanities

Peter Leonard
Yale University Library, Yale University, New Haven, USA.
E-mail address: peter.leonard@yale.edu

Abstract:

This paper considers how libraries can support humanities scholars in working with large digitized collections of cultural material. Although disciplines such as corpus linguistics have already made extensive use of these collections, fields such as literature, history, and cultural studies stand at the threshold of new opportunity.

Libraries can play an important role in helping these scholars make sense of big cultural data. In part, this is because many humanities graduate programs neither consider data skills a prerequisite, nor train their students in data analysis methods. As the 'laboratory for the humanities,' libraries are uniquely suited to host new forms of collaborative exploration of big data by humanists. But in order to do this successfully, libraries must consider three challenges:

1) How to evolve technical infrastructure to support the analysis, not just the presentation, of digitized artifacts.

2) How to work with data that may fall under both copyright and licensing restrictions.

3) How to serve as trusted partners with disciplines that have evolved thoughtful critiques of quantitative and algorithmic methodologies.

Keywords: Digital Humanities, Text Mining, Data Mining, Digital Humanities Librarians.

Introduction

My title at Yale University Library is "Librarian for Digital Humanities Research," and since I am the first person to hold this position, when I arrived last year I had to define exactly who I was and what I would do. Part of this was figuring out what to do on a day-to-day basis, and the other part was how to explain it to others on campus. My talk today comes from the middle – not the beginning, but certainly not the end – of that process. In my presentation I'll

be talking about mining large datasets for the humanities, which is a significant part of my job as I have defined it.

The overwhelming amount of digitized cultural data now available to humanists is a fact of contemporary scholarly life. Some of this originates within the walls of libraries themselves — Yale has digitized hundreds of thousands of pages over the last few decades — some of it comes from nonprofits such HathiTrust, ArtSTOR, and JSTOR, and still more comes from commercial vendors such ProQuest and Gale/Cengage. But regardless of the source, there's an unprecedented amount of text, images, music, and other artifacts of human culture in digital form — arguably more than any one human being could ever read, view, or listen to during their lifetime, even if they restricted themselves to their domain of interest.

Because of this fact, I have tried to define my work at Yale around the question, "How can libraries support humanities scholars in making sense of large digitized collections of cultural material?" Since this has proven to be a very interesting and productive area for me to work in, I want to organize my talk around three specific sections of this question. I'll style them in boldface here:

> How can libraries support
> **humanities scholars**
> in
> **making sense**
> of
> **large digitized collections of cultural material**?

I come from a literature background myself, so I want to begin with the notion of "*humanities scholars*" — what does it mean for humanists to tackle big digital data? The core of my talk, however, will be about the problem of "*making sense*" — what are the tools and approaches that humanities scholars are using to construct meaning from large digital datasets? Finally, I'll discuss some of the problems of these "*large digitized collections of cultural material"* — what are the challenges of working with in-copyright material and licensed material?

"Humanities Scholars"

First, I want to focus on the phrase "humanities scholars." This a tricky term to define precisely, and there are probably more risks to getting it slightly wrong than benefits to getting it exactly right. But to take myself as a safe example: my training in literature involved no classes in statistical analysis, quantitative methods, computer programing, or anything similar. Although the boundaries of the humanities can be fuzzy, it is worthwhile to think about what our common limitations are as we engage in quantitative work – as well as how libraries can address them. What kinds of data mining training and workshops could librarians offer to help meet this emerging need? In addition to the workshops we're developing at Yale, I plan to keep a close watch on my colleagues at Duke, Brown, Columbia, UIUC, and other institutions with Digital Humanities Librarians.

Disciplines outside the humanities, of course, have valuable experience we sometimes lack. Corpus and computational linguists, for example, have been mining large amounts of text for decades — indeed, this is often the core of their domain expertise and professional practice. We shouldn't forget this as humanists struggle to deal with massive datasets on very limited budgets. For every literature graduate student who wants to track changing discourse in historic newspapers, there is almost certainly a parallel graduate student in corpus linguistics

who already knows how. Many digital humanities projects have benefited from pairing students in this way – and what better place for this collaboration than the library? Collaborative research space, such as UCLA's Young Library Research Commons, offers attractive and functional neutral space at heart of campus for these kind of partnerships.

An important caveat to keep in mind, however, is that even though the tools may be shared, the motivations and implications often remain stubbornly bound to separate disciplines (or even individual departments). The lesson for librarians, I believe, is to encourage collaboration when possible, but also to maintain realistic expectations for the limits of cross-disciplinarity in this work.

One of these limits is important enough to warrant special mention. It will come as no surprise to librarians, or to anyone who reads the *Chronicle of Higher Education*, that the humanities have evolved deeply felt and theoretically rigorous critiques of many forms of quantitative analysis. This is linked to what C.P. Snow identified as the divide between the "two cultures," science and the humanities. But when dealing with big data, this abstract issue suddenly becomes a good deal more concrete. How can librarians help humanists with the technical aspects of big data, while respecting their perspectives on evidence, empiricism, and scholarly argument? The support that well-equipped research libraries bring to data science (programmers, experimental design, statistical help) is no doubt useful to students and professors in the humanities. But these scholars may be still coming to terms with the amount of data available to them and the skills required to make sense of it, while simultaneously finding the right way to articulate their interest in this new area. They may even be defending this interest against sometimes-hostile professional peers (if the recent MLA panel "The Dark Side of Digital Humanities" is any indication.) To stay abreast of this conversation, librarians might consider recent books by Jockers, Ramsey, Moretti and other contributors to the emerging discussion about new forms of scholarship.

The foregoing is intended neither to discourage library involvement with digital humanities, nor to cast all humanists as ill-at-ease in the modern, data-centric world. Rather, I mean to encourage involvement in making sense of large digitized collections by *all* relevant members of the library community. This should include data librarians and statistics support staff, but also the humanities librarians with whom I share an office at Yale. Subject librarians in English, history, and other humanistic disciplines broaden the skills of a team, because knowing how material is interpreted and used in disciplinary conversation is at least as valuable as knowing an algorithm to process it. Bringing the full spectrum of library team members to bear on these projects can increase the chances of projects being successful.

"Making Sense"

After that initial admonition to involve many kinds of librarians in these efforts, the question of how we can together help humanities scholars "make sense" of digital material forms the core of my talk, and is the central challenge I address in my job. In the past, digital library systems, as well as the presentation layers of commercial vendors, have focused on the *display* of material, rather than its *analysis*. In the best of situations, libraries brought to bear their cataloging and metadata expertise to shape online collections into interpretable chunks, but this was not always possible with large and highly-granular digitization projects. (Box 12, Folder 7 only goes so far as an organizational facet.)

This approach is no longer adequate — there is simply too much digital cultural material online, event at smaller institutions and even in small fields. Now is the time for libraries to

build on their display and metadata expertise, by pushing further to develop and implement systems that *make sense* of digital cultural collections. This does mean an investment in new kinds of infrastructure, but it does not mean starting from scratch or reinventing the wheel.

I want to suggest two types of sense-making that libraries can apply to large digitized collections of cultural material, each grounded in a radically different perspective. For each of these two approaches, I'll show some tools that I think are good examples of the assumptions, benefits, and limitations of the approach. The two approaches can be summarized as follows:

1) Looking for something you think is there

2) Letting the data organize itself

The first of these will seem similar to current practice in the humanities, while the second may seem anathema to it. (Can data have agency?) But in addition to explaining what I mean by letting data organize itself, I hope to suggest some ways in which these two approaches may profitably be combined. First I'll detail some examples of the former: looking for something you know is there.

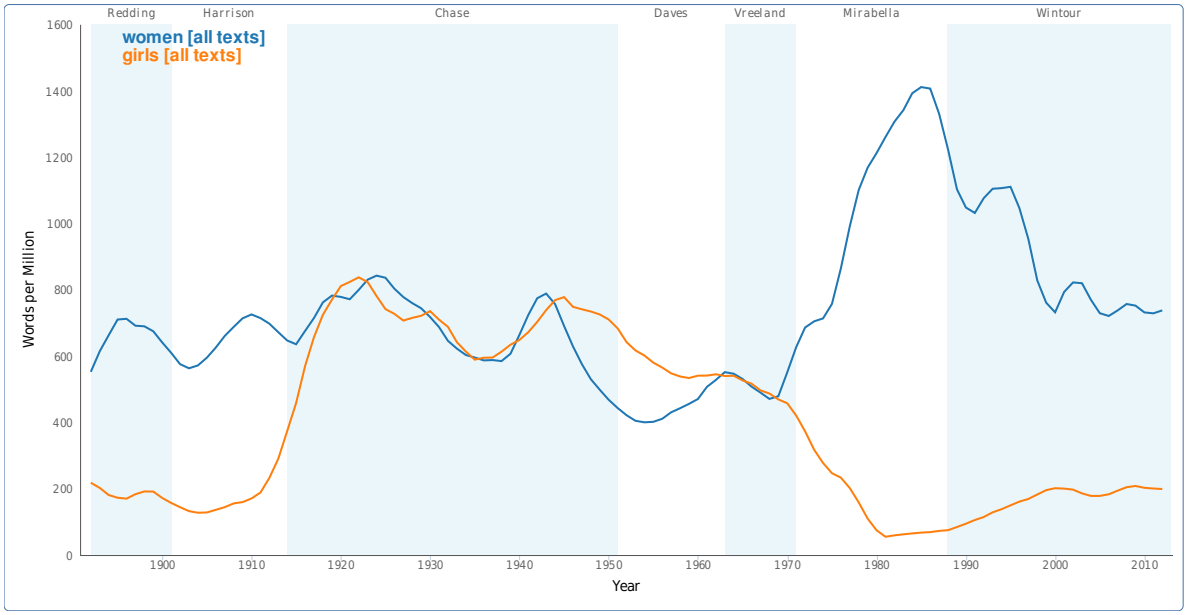
Perhaps no other project is embedded in the public imagination as an exemplar of "digital humanities" as the Google Books N-Gram Search.[1] Launched in 2010, this online tool was a joint venture between Google and scholars at the Harvard Cultural Observatory. It allowed users to look for words and phrases in the vast majority of volumes in Google Books — including, crucially, those volumes that were still in copyright. By only allowing users to chart the rise and fall of n-grams (unigrams, bigrams, trigrams, etc), the tool neatly sidestepped the problem of full-text access to copyright-protected volumes. The result was a narrowly-focused tool, providing analysis of current and historical texts, that captured the imagination of the broad public as well as scholars.

It turns out, however, that as compelling as 13 million books are, some research questions are best answered by *smaller* corpora, not larger. One of the most exciting advances in the years since the launch of the n-gram tool has been the development of an open-source implementation, created by some of the same individuals, that can be brought to bear on collections of texts. Dubbed *Bookworm*, this independent "bring-your-own-books" tool has already been deployed on subject matter as diverse as the *arXiv* pre-print repository in the sciences,[2] the texts of laws passed by the US Congress,[3] and historical newspapers in America.[4]

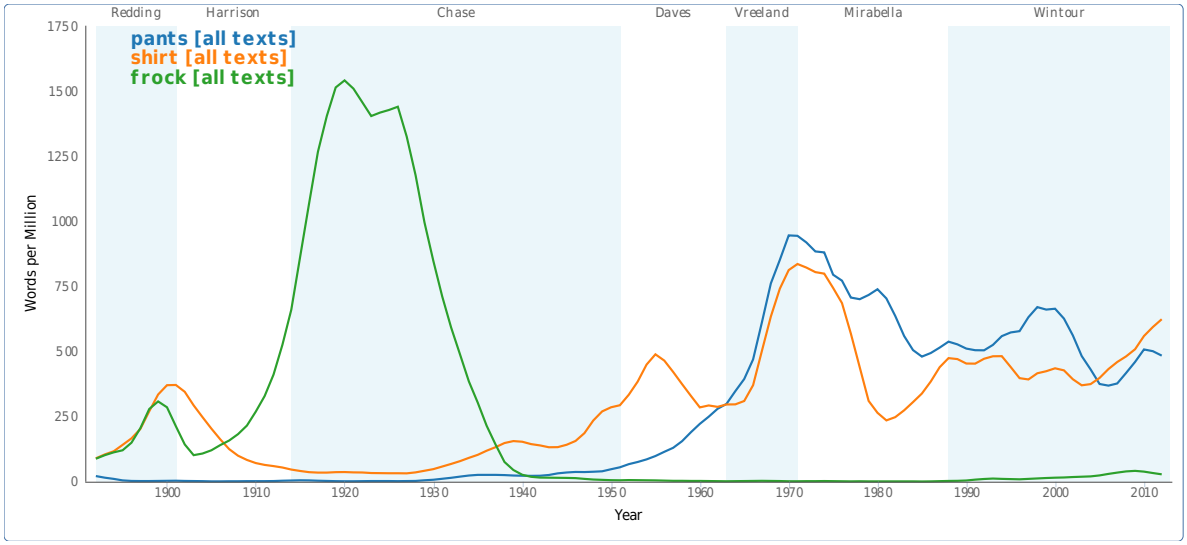
At Yale Library, we were struck by the possibility of using this tool to help researchers make sense of our large text collections — both those we have created ourselves and those we license from vendors. Bookworm is open-source, and built upon industry standard software layers such as Python and MySQL, so it's relatively easy to deploy for testing purposes. We have a number of demonstration projects underway, but the one I want to mention here is an example of this idea of a smaller, focused corpus. It's a collection of hundreds of thousands of articles and advertisements from 1892 to the present day: the *Vogue* archive, digitized for

[1] See Cohen 2010
[2] http://bookworm.culturomics.org/arxiv/
[3] http://bookworm.culturomics.org/congress/
[4] http://bookworm.culturomics.org/ChronAm/

Condé Nast by ProQuest. ProQuest already maintains a digital platform for subscribers to the *Vogue* archive to browse through the 400,000 pages in the collection, and we had no interest in duplicating that display system. Instead, we wanted to experiment with an analysis tool — using the Bookworm n-gram search tool to help identify both continuity and change in the archive. Two ideas motivated us — first, that *Vogue* is not only about fashion. Second, fashion is, almost by definition, subject to temporal change. Here are two examples of n-gram searches that show this.



This first example draws upon two collective nouns that we suspected might diverge over time: **girls** and **women**. Ben Schmidt from Northeastern University, one of the co-authors of Bookworm, suggested graphing their relative frequency. As evident from the graph above, the most striking pattern from 1892 to 2013 is the apparent decline of the word **girls** and the rise of the word **women** at the beginning of the 1970s. The alternating vertical bars represent the changing editors-in-chief of *Vogue* over time, and this rise in **women** at the expense of **girls** coincides with the ascension of Grace Mirabella in 1971. Whether due to a change in values imposed from the top, or to a shift in how society understood those terms more generally, our supposition that these two terms would diverge was confirmed by this graph of their relative frequency. This is an example of how *Vogue* is not only about fashion. The peaks of the graph point to where a social historian might focus attention – a call to action for close reading, rather than a deterministic answer in itself.



My second example draws upon three terms for items of clothing: **pants**, **skirt**, and **frock**. We chose these terms because we suspected they might show historical variation, and indeed, the graph of their relative frequency shows precisely that: **frock** peaks in the 1920s, before essentially falling out of use other than a very small revival around 2010. **Pants** starts trending up around 1950, peaking at the same moment in the early 1970s associated with the switch from **girls** to **women** noted above. **Skirt** demonstrates the most continuity, never falling below a certain threshold and essentially holding steady with **pants** ever since the 1970s.

At this point you can probably imagine all sorts of terms you'd like to graph yourself, and that's the point of the Bookworm tool — by ingesting and indexing millions of words and two-word phrases, it allows you to perform experiments on large collections of text almost instantaneously. This is an excellent example of libraries building infrastructure for sense-making, instead of merely display. Libraries should be using Bookworm, and tools like it, to transform passive archives of digital text into active sites of research and direct engagement. And the utility of a tool such as this does not cease when an article goes to publication: scholars can footnote a deep link to a dynamic graph so that users can adjust parameters (smoothing, case-sensitivity, etc.) at will.

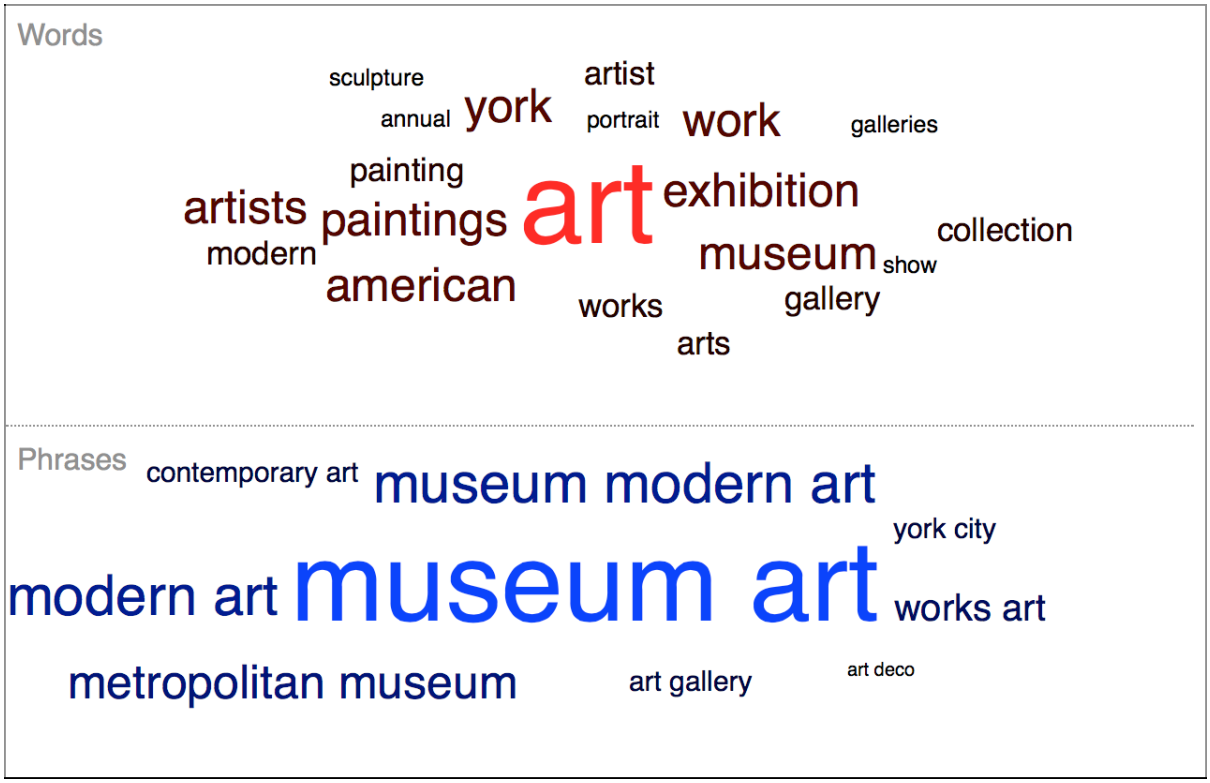
The above examples depend on a partnership between a software tool and a human to think of words to look for. It takes a person with at least a passing knowledge of Western fashion history to know that **frock** might be a productive term to measure against **pants** and **skirt**. But what about those constellations of phrases and words we don't yet know about? After all, the only way to know a corpus completely is to read it thoroughly to glean the other terms that we should put into our query. But at over 400,000 pages, the complete archive of *Vogue* magazine would test the patience of even the most devoted fashion maven. We are faced with the paradox of the Borgesian map, frustrated by the impossible task of constructing a mental model of the archive at one-to-one scale.

Luckily, there are domains outside the humanities that have confronted the problem of unlabeled texts — that is to say, documents about which we don't know very much in advance. Computer science, mathematics, statistics, and information retrieval have all contributed to a complex set of theories and approaches which, in their current state of development, promise to do something on the edge of witchcraft: let data organize itself. The

technique I'm going to talk about here, Topic Modeling, is one that extracts latent themes from a large corpus of text, by means of a computer program that does not understand English and knows nothing about the time period or cultural context in which those texts were written.

It may disappoint us to know that this magic trick is predicated on the simplest of methods: term co-occurrence. In the words of the British linguist John Firth, "You shall know a word by the company it keeps." (179) It's famously difficult to explicate the mathematics behind modern Topic Modeling (which involves such concepts as Latent Dirichlet Allocation and Gibbs Sampling), but what is relevant here is that Topic Modeling discovers statistical patterns of words that occur near each other and — with some help fine-tuning from a human with a basic understanding of the corpus — it can produce uncannily interesting results.

As always, it is better to show than tell. Let me give you two examples from the same corpus, *Vogue* magazine. We examined 100,000 articles and asked the Topic Modeling algorithm to produce a list of twenty topics that characterize the discourse therein. Here is one of these topics, expressed as a word cloud of the most significant unigrams and bigrams (one- and two-word phrases) in the discourse. Because humans, rather than computers, are responsible for assigning a label to each topic, I'll refer to this by its original name of **Topic 1:**



You'll probably have no difficulty divining the label we gave this topic: art and museums. It may surprise you to know the depth of coverage *Vogue* has given the art world over the years. Although the data we received from ProQuest was well marked-up in terms of author, title, photographer, and advertiser, it did not contain any thematic categories or subject headings. Because of this, discovering this art discourse through title search alone would have been difficult: neither "art" nor "museum" occurs in the title of "de Kooning and the Old Masters." However, the Topic Modeling algorithm noticed that the words on this slide

co-occurred with one another at a significant rate, and essentially created a virtual subject heading out of thin air.

Continuing the metaphor of a subject heading, the topic modeling algorithm can tell us which articles are the most "saturated" with a given topic[5] — in other words, which articles have the highest concentration of the particular pattern of term co-occurrence for a given topic. We can then examine those texts more closely:

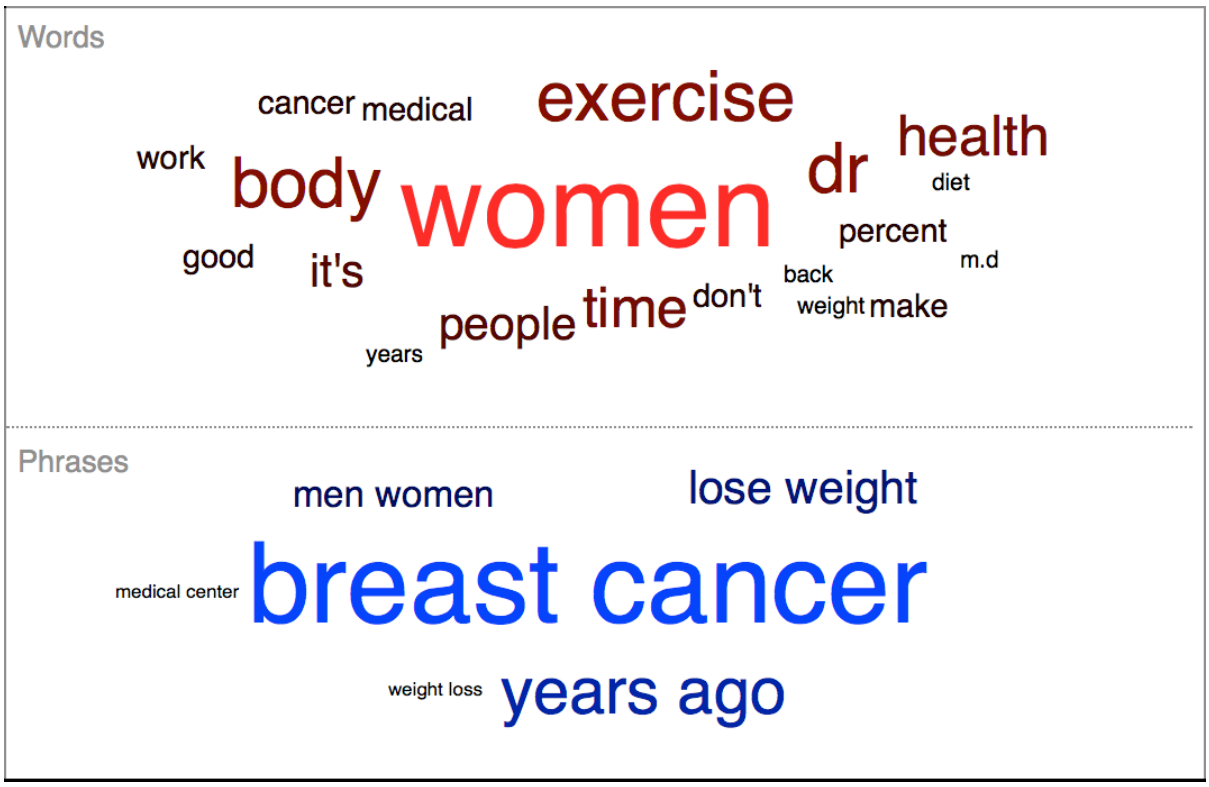
1949

- 96% "People and Ideas: Vogue...A Prize Winner" *Vogue* 1 Jun. 1949: 36
- 77% "People and Ideas: Van Gogh Landmarks" *Vogue* 15 Oct. 1949: 90
- 74% Edward Steichen. "People and Ideas: Irving Penn..." *Vogue* 1 Dec. 1949: 91
- 70% "People and Ideas: Picasso Plate" *Vogue* 1 Jan. 1949: 108
- 64% Saul Steinberg. "People and Ideas: Man Up a Tree" *Vogue* 1 Jul. 1949: 64
- 64% "People and Ideas: Madonna and Child by Henri Matisse" *Vogue* 1 Dec. 1949: 84
- 59% "People and Ideas: Barcelona and Picasso" *Vogue* 1 Jul. 1949: 42
- 58% "People and Ideas: The Eye of American Realists" *Vogue* 1 Feb. 1949: 178
- 57% "People and Ideas: Braque" *Vogue* 1 Apr. 1949: 140
- 55% Dorothy Norman. "People and Ideas: The Great Barnes Collection" *Vogue* 15 Aug. 1949: 128
- 50% "People and Ideas: Elegance and Art" *Vogue* 1 Nov. 1949: 112
- 49% "People and Ideas: Mrs. Dean Acheson" *Vogue* 15 Mar. 1949: 100
- 45% "People and Ideas: Fashion...an Art in the Museums" *Vogue* 1 Feb. 1949: 211
- 44% "People and Ideas: The Modern House of Mr. and Mrs. William Goetz in Hollywood" *Vogue* 15 Apr. 1949: 94
- 43% "People and Ideas: Father Couturier" *Vogue* 15 Feb. 1949: 79
- 42% "People and Ideas: Mrs. Huttleston Rogers" *Vogue* 1 Feb. 1949: 210
- 41% "People and Ideas: French in New York: Balthus, the Painter/Barrault, the Actor" *Vogue* 15 Mar. 1949: 102
- 38% Ione Robinson. "People and Ideas: A Visit to Braque" *Vogue* 1 Apr. 1949: 191
- 37% Rosamond Bernier. "People and Ideas: Matisse Designs a New Church" *Vogue* 15 Feb. 1949: 76

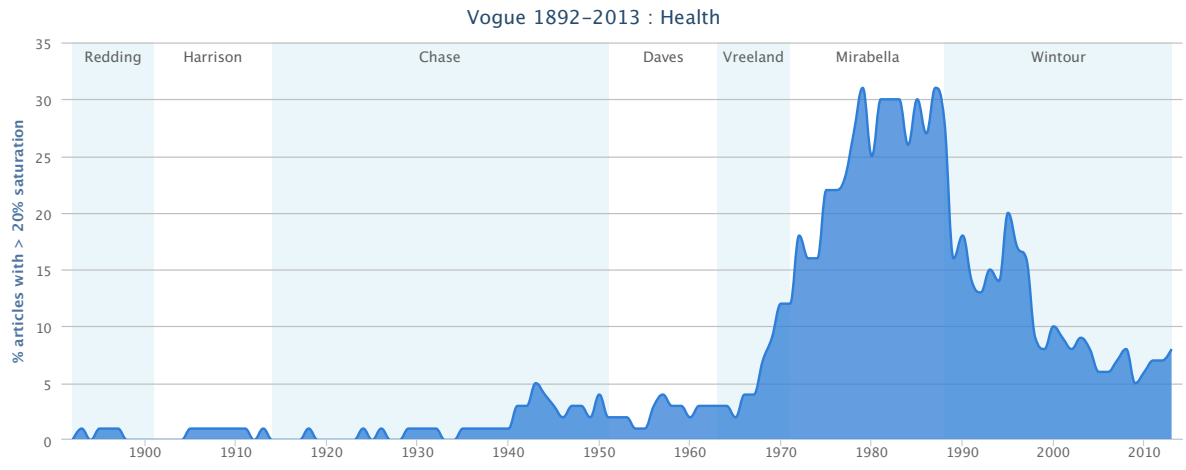
In this way, a text-mining technique with origins in computer science actually enables *more* close reading — by pointing to the needles in the haystack in the form of articles saturated with a particular discourse.

My next topic modeling example uses this "saturation" data to show how we can measure the rise and fall of these discourses over time. This example centers on a discourse that the computer generated as "Topic 14":

[5] An article can participate in many discourses simultaneously, which is an advantage of Topic Modeling over other machine learning techniques such as Classification or Clustering.



Most of us would immediately intuit that the words and phrases shown above are actually characteristic of **women's health**. This another somewhat surprising topic for a magazine dedicated to *haute couture*. Would we have thought to look for "breast cancer" in the Bookworm tool? Perhaps not — and so we turn to another facet of the data to try and understand this topic: the diachronic axis. Charting the percentage of articles saturated with this topic over time shows the following pattern:



The editorship of Grace Mirabella brought about a sharp increase in articles concerning women's health, at least according to the 20-topic model we produced. There are two interesting ways to explore this graph interactively, which I can do by clicking on a given year along the x-axis. Firstly, we can see the articles at the zenith of the topic's popularity:

99% "Q & A: The Pill" 1 Dec. 1987: 361
98% Jane Ogle. "Facts on Fat: Obesity—a Heavy Health-risk Factor" 1 Aug. 1979: 249
96% Melva Weber. "Inner Info: Contraception" 1 Aug. 1978: 210

95% Charles Kuntzleman. “What Is the Best Way That You Can Shape Up for Active Sports?” 1 Aug. 1979: 82
95% Jane Ogle. “Why Crash Diets Don't Work” 1 Aug. 1979: 248
91% Melva Weber. “Latest in the IUD Dust-Up...” 1 Mar. 1975: 88
89% Ellen Switzer. “Your Blood Pressure” 1 May. 1973: 152

Second, we can examine years many decades before the 1970s, for traces of this topic long before it rose in importance:

66% “Correct Breathing as a Figure Builder” 13 May. 1909: 894
50% “How to Reduce Weight Judiciously” 15 Jun. 1910: 10
44% “Health Laws for Rheumatics” 15 Mar. 1911: 100
43% “Mechanical Massage” 18 Jul. 1907: 84
29% “Teaching Poise to Children” 11 Sep. 1909: 342
26% “Tuberculosis: A Preventable and Curable Disease” 12 Aug. 1909: 188
26% “Good Form for These Ruthless New Dresses” 15 Apr. 1931: 93

These two examples of the topic over time suggest both that the discourse was present during most of the magazine’s history — and also that it greatly increased in importance under Mirabella in the 1970s. According to my colleague Lindsay King in the Arts Library, Mirabella’s goal as editor, stated in her memoir and numerous interviews, was to make *Vogue* more relevant to all aspects of the modern working woman’s life. (She was also married to a physician.)

Important to note is that the Topic Modeling approach is superficially similar to, but fundamentally different than, the N-gram Search approach. As expressed in this project, both techniques result in a rising or falling graph over time. But Topic Modeling is not merely measuring the presence of the few words I gave above as indicative of the discourse. Rather, a topic consists of a probability distribution over all words in *Vogue*. As such, a topic is an extremely powerful tool that can capture a shifting discourse over time, even when individual terms come and go. But it can’t satisfy every need - certain words and phrases will be so important that n-gram search is a better approach.

This point leads me to my conclusion of this section, which is to make an argument for both kinds of scholarly interrogation of digital archives represented by two very different approaches:

1) Looking for something you think is there

2) Letting the data organize itself

By building both kinds of tools on top of digital collections, libraries (and librarians) can help humanities scholars evaluate many different kinds of technologically-enabled research methods, without individual professors and students being forced to invent their own software and set up their own servers.

I hope that sense-making tools such as Topic Modeling and Bookworm will be an increasingly important part of the work we do at Yale Library — and not only on vendor collections such as *Vogue*. We have plans to deploy both tools on an internally-digitized collection, *Yale Daily News*, which has been published continuously since 1878 and is thus the oldest college daily newspaper in the United States. Other examples of these tools’ uses include subsets of journal articles downloaded from JSTOR, or a set of the volumes from the HathiTrust Research Center.

"Large Digitized Collections of Cultural Material"
This diversity of applicable datasets leads me to the final point I want to address: the peculiarities of working with large digitized collections of cultural material. Many of you will be familiar with these organizations:

- HathiTrust
- Getty
- JSTOR
- Internet Archive

But alongside these public-facing websites and brands are some less well-known research projects, with special importance for data mining:

- HathiTrust **Research Center**
- Getty **Research Portal**
- JSTOR **Data for Research Program**
- Internet Archive **Bulk Download Advanced Search Tool**[6]

These have been set up to facilitate large-scale data analysis for their respective collections. Alongside technical affordances such as bulk downloads, you'll also find policy declarations, guidelines for use, and other documents of interest to libraries and scholars alike. Crucially, these non-profits have been at the forefront of thinking about how to make in-copyright material available for Text and Data Mining (TDM). JSTOR's Data for Research program (DFR), for example, makes term 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. I expect HathiTrust's eventual strategy for researcher access to parts of the Google Books corpus to look similar to DFR.

Moving beyond the non-profit archives and research centers listed above, there is increasing interest in working with digital data provided by commercial vendors such as ProQuest and Gale/Cengage. An early call to action about the potential for these datasets was a presentation given by Duke University's Joel Herndon and Molly Tamarkin at the 2012 Coalition for Networked Information. Their talk on "What To Do With All Of Those Hard Drives" (with accompanying slides available online) is a wonderful introduction to the possibilities of working with "local copies" of vendor data. This means the physical hard drives that many research libraries have been accumulating over the years, which contain duplicates of content for which libraries have perpetual access licenses.

Of course whenever engaging with commercially-licensed material, consulting with a licensing team, copyright librarians and legal counsel is a must. But in working with vendor-supplied collections of cultural material, I've found it useful to break down the problem of data mining into two domains: *analysis* and *presentation*.

Librarians, and the scholars they serve, need full and unfettered access to the raw digital data for the purpose of *analysis*. This is consistent with the December 2013 IFLA statement on Text and Data Mining (TDM), which notes:

[6] See Internet Archive 2012

> As TDM simply employs computers to "read" material and extract facts one already has the right as a human to read and extract facts from, it is difficult to see how the technical copying by a computer can be used to justify copyright and database laws regulating this activity. (IFLA 2013)

We cannot perform data mining by using existing web-based presentation layers, because downloading the entirety of the data is either impractical (the *Vogue* dataset is six terabytes) or likely to result in automatic restrictions such as the Captcha system deployed by ProQuest in August 2013.[7]

However, after Text and Data Mining has been performed, vendors' existing web-based systems have an important role to play in the *presentation* of the data being mined. Most concrete technical implementations of what Franco Moretti has termed *distant reading* let scholars "zoom in" to the granular unit — a paragraph, sentence or word — of analysis. The Bookworm n-gram search tool I showed previously is an example of how this works in practice: clicking on any point in the trend line shows a list of permalinks (stable URLs) into the ProQuest "system of record":

This implementation allows us to do the analysis work that we consider our core interest (making sense of data) while freeing us from the burden of re-presenting all of the texts and images in a vendor collection. It is a concrete instantiation of the value espoused in the 2013 IFLA TDM statement that "Researchers must be able to share the results of text and data mining, as long as these results are not substitutable for the original copyright work." In addition, there are two further benefits to consider.

First, having the representation of the articles remain on the ProQuest site implicitly validates the central role of the vendor in maintaining this information in exchange for a licensing fee. The links that are built in to our systems will fail to resolve unless the user is on a university network whose library subscribes to the electronic resource — and that is entirely by design. It's even possible to argue that as compelling Digital Humanities work gets done on top of systems that point back to vendor websites, the demand for these digitized archives will increase as scholars petition their libraries to subscribe.

Second, when one of our users investigates a deep link that's emblematic of a larger pattern,

[7] See Sidi 2014

there's a chance that he or she will browse "horizontally" through the vendor's system. Come for the article about skirt length, stay for the articles about women's professional dress, so to speak. Libraries are paying good money for access to these electronic resources, and many of us suspect they are under-utilized in proportion to their cost. While browsing through 400,000 magazine pages may be overwhelming, entering through a specific point in the archive identified by an algorithm solves the "where do I start?" problem, and can lead to serendipitous exploration of the database itself. Easing researcher access to information has been at the core of most modern libraries' value systems, and Text and Data Mining is an exciting new way to use Big Data methods to facilitate the kind of targeted close reading most humanists are well trained in.

Conclusion

Mining large datasets for the humanities is a nascent field, and thus full of uncertainties. Humanists, and humanities librarians, need to learn to work alongside other disciplines in new ways, experiment with novel ways of sense-making, and deal with unfamiliar restrictions on digital material. This talk represents my imperfect and early understanding of some answers to those uncertainties that are developing at Yale. I look forward to a conversation about how we can work together as librarians to address these challenges. If successful, libraries stand to profit in two ways. First, they can extend their presence further into the research workflow by helping to *make sense* of digital data, along with acquiring and preserving it. And second, they can assure better use of their licensed electronic materials by scholars – materials which are increasingly forming the lions' share of many research libraries' collection budgets.

Acknowledgments

McCallum, Andrew. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

Schmidt, Ben; Camacho, Martin; Cherniavsky, Neva; Lieberman Aiden, Erez; and Michel, Jean-Baptiste. *Bookworm*. http://bookworm.culturomics.org. 2013.

Yale Library: Daniel Dollar, Michael Dula, Joan Emmet, Lindsay King, Julie Linden, Alan Solomon.

References

Aiden, Erez, and Jean-Baptiste Michel. *Uncharted: Big Data as a Lens on Human Culture*. 2013. Print.

Cohen, Patricia. "In 500 Billion Words, a New Window on Culture." *The New York Times* 16 Dec. 2010. *NYTimes.com*. Web. 28 May 2014.

Firth, J. R. *Papers in Linguistics, 1934-1951*. London; New York: Oxford University Press, 1957. Print.

Grusin, Richard. (Moderator) "The Dark Side of Digital Humanities." Modern Language Association Convention. Boston. 2013.

Herndon, Joel, and Molly Tamarkin. "What to Do with all of those Hard Drives: Data Mining at Duke." Coalition for Networked Information. Washington DC. 2012. http://www.cni.org/topics/digital-libraries/hard-drives-data-mining-duke/attachment/cni_what_tamarkin/

IFLA. "IFLA Statement on Text and Data Mining." N. p., 19 Dec. 2013. Web. 28 May 2014.

Internet Archive. "Downloading in Bulk Using Wget." 26 Apr. 2012. Web. 28 May 2014.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. Urbana; Chicago; Springfield: University of Illinois Press, 2013. Print.

Mirabella, Grace. *In and Out of Vogue*. 1st edition. New York: Doubleday, 1995. Print

Moretti, Franco. *Distant Reading*. 1 edition. London ; New York: Verso, 2013. Print.
Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press, 2011.
Sidi, Rafael. "ProQuest Service Alert." *ProQuest Blog*. N. p., 28 Jan. 2014. Web. 28 May 2014.