The Remaking of Reading: Data Mining and the Digital Humanities

Matthew G. Kirschenbaum
University of Maryland
mgk@umd.edu

Abstract

This paper discusses applications of data mining in the seemingly unlikely field of literary criticism. While the underlying techniques are traditional—Naïve Bayes, SVM—literary criticism, and the "digital humanities" more generally, differ from other domains in that they rarely admit ground truth into their discussions. Instead, data mining and machine learning are best understood in terms of "provocation"—the potential for outlier results to surprise a reader into attending to some aspect of a text not previously deemed significant—as well as "not-reading" or "distant reading," the automated search for patterns across a much wider corpus than could be read and assimilated via traditional humanistic methods of "close reading." At a moment when a widely publicized report by the National Endowment for the Arts concluded reading itself was "at risk," large online text collections (Google Books, the Open Content Alliance) are making millions of texts available in machine-readable form. Data mining is part of this remaking of reading.

1. Introduction: The Digital Humanities

The humanities encompass what are popularly known as the liberal arts: literature, history, art history, philosophy, music, and language studies. Computational techniques in the humanities have a long history, dating back at least as far as Father Roberto Busa's use of punch-cards and an IBM computer to compile a concordance to the complete works of Thomas Acquinas in the late 1940s [8]. "Humanities computing," as it since came to be known, developed alongside of corpus linguistics and included applications for stylistics, stylometrics, and author attribution. In the 1980s, practitioners embraced SGML and developed DTDs suitable for rich text representation (notably the Text Encoding Initiative, or TEI), introducing tools and software for a wide range of different kinds of textual analysis. Simultaneously, the advent of first desktop multimedia and then the Web led to increased interest in the digital presentation of non-textual forms, music and the visual arts. Work in the 1990s was thus characterized by extensive text and image archives, or "thematic research collections," often including digital facsimile representations of multiple versions of manuscripts and other rare and precious material, all fully searchable via machine-readable markup. More recently still, the field—increasingly known as simply "digital humanities," the shift from an adjective and a gerund to a noun phrase perhaps indicative of its growing self-confidence and relevance—has turned to data mining as offering potentially powerful new methods for finding patterns across large text collections [7].

The wide-spread availability of large electronic text collections, coupled with the promise of metadata and text visualization offers the means for bringing these tools within reach of even the "lay-humanist," that is the traditional scholar of arts and letters not well-versed in computational method. At the same time, it must be understood and acknowledged that there is a deep tradition of skepticism towards quantitative and empirical techniques among humanists, which too often smack of positivism and objectivity in domains for which interpretation, ambiguity, and argumentation are prized far above ground truth and definitive conclusions. The humanities embrace a culture of conversation, not problem-solving. Humanists do not seek to address the "problem" of Emily Dickinson in order to move on to the more vexing problem of Walt Whitman. Any serious adaptation of computational tools in the humanities must acknowledge the limitations of regularized methods in the face of poetry, philosophy, and the rich variety of human thought and expression.

2. Close Reading, Not-Reading, Distant Reading

Whatever else they might do, humanities scholars read. In some fields, like literature, the text is both primary and secondary, since it is both the object of study and the vehicle for scholarly communication. In other fields, like art history or musicology, the objects of study may primarily be non-textual in form, but text is still the essential conduit for scholarly communication in the guise of catalogues, treatises, monographs, articles, scores, libretti, and so forth. The reading of written text is thus essential to humanistic study.

But what is reading? We do not, after all, read a novel the same way we read a reference work. We do not even

read a novel the same way we read a poem; for a piece of verse by William Carlos Williams, a scholar will linger over every word, even its physical placement on the page. This is not the way we typically read *War and Peace*. Some books are read for immersion—the familiar and comfortable image of a reader under a tree or in some other naturalistic setting is a reflection of the "silence and slow time" we typically associate with such reading, that is the deep, meditative pleasure of becoming "lost" in a book. But not all, indeed not most, books are destined for this kind of reading. Pictures of medieval saints at their desks reveal surprisingly complex scenes of reading, with specialized devices and furniture for holding multiple books open at once, the better to allow a reader to cross-reference and perform look-up tasks. Thomas Jefferson's famous revolving bookstand, which allowed him to keep five volumes splayed open and available within the visual field of his reading, is an Enlightenment refinement of these same reading technologies. Books are random access devices par excellence, and the strict linear sequences of reading we associate with sitting under the tree is the exception, not the rule [3].

Once we acknowledge that there are different kinds of reading, we begin to see the value in techniques that have lately become known as "not-reading" or "distant reading." To quote Martin Mueller:

> As long as there have been books there have been more books than you could read. In the life of a professional or scholar, reading in the strong sense of "close reading" almost certainly takes a back-seat to finding out what is in a book without actually reading all or even any of it. There are age-old techniques for doing this, some more respectable than others, and they include skimming or eyeballing the text, reading a bibliography or following what somebody else says or writes about it. Knowing how to "not-read" is just as important as knowing how to read [5].

Distant reading (or "distance reading"), meanwhile, is the coinage of Franco Moretti, the Stanford literary scholar who has long been an advocate for using statistical, quantitative methods to "read" large volumes of text at a distance, using "graphs, maps, and trees" as forms of abstract representation that enable the study of patterns over time. A typical problem for Moretti might ask what we can learn about the history of the novel by studying data about publication trends for hundreds of novels over the course of a century [4].

It will come as no surprise that the programmed abilities of digital tools are compatible with such precepts. The adoption of computational techniques within the humanities allows us to build tools that support the basic tenants of not-reading or distant reading as described here. The significance of those terms is not in their novelty, but rather precisely the way in which they are able to draw connections between cutting edge technologies and long-recognized patterns of human behavior when engaging in the indispensable activity of reading.

3. Mass Book Scanning

The number of books in the world, while very great and increasing daily (indeed, hourly), is not infinite. By most estimates there are 50-60 million books in the world. The Library of Congress holds some 32 million volumes on its shelves. The British Library, 25 million. The National Library of China, 10 million. We can thus reliably speak of the number of books in the world as measurable in the tens of millions [11].

Recently, mass-scale book scanning projects have promised to make good on what were once utopian dreams of digitizing all of the world's book-based knowledge [1]. The most highly-publicized and ambitious of these is Google Books, which, in partnership with a number of the world's leading research libraries, aims to scan 15 million books within the next decade. While much will still remain unscanned, the Google Books project (and others like it, such as those by Microsoft or the non-commercial Open Content Alliance and Project Gutenberg) promises to make access to textual information available on an unprecedented level. Even today we see the impact of mass book scanning in systems like Amazon.com, where a statistical profile of the text is presented alongside of sales information (allowing a user to "not-read" it according to the Flesch-Kincaid index, which measures "readability," or compare the distribution of number of syllables per word or number of words per sentence as an indicator of "complexity"). More profound, of course, is the ability to search large machine-readable collections for key words and phrases, with results delivered to the user either as "snippets" of the key word in context or (copyright permitting) as full-text access. The ability to annotate and leave comments on individual electronic books is likewise already in place, and thereby lays the groundwork for a secondary layer of textual information which, in Talmudic fashion, is itself indexable, searchable, and mineable.

Numerous smaller, more specialized collections also exist, and these are likely to be of even more immediate value to scholars. For example the Wright Fiction Archive, a digital collection of every novel published in America between 1851 and 1875, some 2800 works in all; access to such a corpus—and the tools with which to not-read or distance read several thousand novels published within a decade of the beginning and end of the Civil War—suggests important new horizons for literary scholarship.

Mass book scanning is not perfect, and some of the issues and challenges are obvious. Licensing, copyright, and intellectual property may prove to have undue influence over what is and is not included in the world's stores of digitized books. The quality of the scanning has been brought into question, as have choices about which editions get scanned. And then there are the intangible aspects of books which do not translate well to the screen—the heft of a volume, its look and feel, which contribute to the reader's experience of the text. We do not read a letterpress volume in a hand-tooled leather binding in quite the same way we do a mass-market paperback; yet these differences, as well as the characteristics of individual copies—a dog-eared page which serves to mark an important place in the text—tend to become flattened and erased in the digital display of the book. Nonetheless, while we may debate whether or not books will ever be replaced by their digital surrogates, it is clear that the existence of collections of millions of books in machine-readable form will have its impact on humanistic studies, supplementing if not replacing the libraries of individuals and institutions. As such collections become available, we will develop the tools with which to not-read with them.

4. Provocation: The Nora Project

Nora (www.noraproject.org), named for a character in the William Gibson novel *Pattern Recognition*, is a classification and prediction system comprising an OpenLaszlo Web interface built on top of the D2K (Data2Knowledge) toolkit by the Automated Learning Group at UIUC [6]. In its current configuration, Nora permits the user to choose one of three text collections (non-fiction materials from Documenting the American South, several hundred poems and letters by Emily Dickinson, or a small set of sentimental novels), and perform a classification exercise using either Naïve Bayes or Scalable Vector Machines (SVM). The assumption is that a user comes to Nora interested in testing some hypothesis that is tractable to classification and prediction; for example, the user might be interested in the characteristics of erotic language in Dickinson's poetry (a well-turned question in the scholarship). The goal is not to use the machine to supplant the judgment and expertise of a human expert who has spent a lifetime reading Dickinson, but rather to see if the classifications can "provoke" new insight amongst a body of familiar texts. This point is particularly important since scholars know that reading and rereading is essential to the process of interpreting a complex author like Dickinson, but with repeated reading—this is an unavoidable feature of human cognition—complacency, even tedium inevitably sets in. The process of "training" the classification, coupled with attention to its outcome, is one strategy for overcoming complacency, even in an experienced reader.

Having selected a workset and classification technique, the scholar proceeds to "train" the classifier by first rating some modest number of works on an arbitrary five-point scale, corresponding to greater or lesser subjective determinations of (in this case) eroticism. (Inevitably, we came to call this particular exercise "hot or not.") The classification is then initiated, and the scholar will be presented with the results of the prediction. At this point she may choose to rate additional texts, or adjust the ratings on those returned by the prediction, in order to train the system. This process may iterate until the scholar tires or sufficient insight is judged to have been obtained.

How effective is a system like Nora amongst practicing literary critics? This long quotation captures one scholar's experience, someone who has spent a career as an Emily Dickinson authority:

> When Bei sent the computationally-generated list of found erotic terms and "Vinnie" was a "hot" term, and one of the most frequent to occur, I was at first surprised. But just a smidgeon of reflection changed that surprise to "uh, duh" recognition. Of course I had known that many of Dickinson's effusive expressions to Susan were penned in her early years (written when a twenty-something) when her letters were long, clearly prose, and chock-full of the daily details of life in the Dickinson household. But I had never thought of this fact in quite the way that the data mining "search and find the erotic" exercise made me put together the blending of the erotic with the domestic. And thus I was surprised again because I've written extensively on the blending of the erotic with the domestic, of the familial with the erotic, and so forth. So I should have expected "Vinnie" to appear frequently in these early letters and to appear near erotic expressions, but I was still taxonomizing (and rather rigidly so) in my interpretations without realizing I was doing so. In other words, I was dividing epistolary subjects within the same letter, sometimes within a sentence or two of one another, into completely separate categories, and I was doing so un-self-consciously. I could wax eloquent here about why understanding the erotic as part and parcel of, and not separate from, daily life is so important, but in the interest of time I'll just note the important connection, a connection discouraged by the traditional hierarchies of Western culture. Making the connection leads to critical understandings not otherwise obtainable, and the data mining exercise helped me do that. Similarly, though I had not designated "mine" as a hot word, it did not surprise me at all that it was FIRST on Bei's list. The minute I saw it, I had one of those "I knew that" moments. Besides possessiveness, "mine" connotes delving deep,

> plumbing, penetrating--all things we associate with the erotic at one point or another. And Emily Dickinson was, by her own accounting and metaphor, a diver who relished going for the pearls. So "mine" should have been identified as a "likely hot" word, but has not been, oddly enough, in the extensive literature on Dickinson's desires. . . . So the data mining has made me plumb much more deeply into little four- and five-letter words, the function of which I thought I was already sure, and has also enabled me to expand and deepen some critical connections I've been making for the last 20 years [9].

There are several points worth remarking here. The first is that the machine has not replaced the judgment, insight, and instincts of the human subject expert. Second, while the machine learning algorithms are here iterating over several hundred texts, "reading" them at a distance, the end result is attention to individual words, the building blocks of language and poetry. Given that there is no ground truth is a discipline like literary criticism, it is difficult to know how influential these results will prove. A scholar would have to write them up in traditional article or monograph form, wait for the article or monograph to move through the peer-review process (this can take months or years) and then other scholars in the field will have to read it, be influenced by its arguments, and adjust their own interpretations of Dickinson—in turn publishing these in their own articles and monographs. Nonetheless, we believe that the Nora system has suggested that classification and prediction can be useful agents of provocation in humanistic study.

5. Not-Reading *The Making of Americans*

The Making of Americans by Gertrude Stein (1925) is an experimental novel representative of literary "modernism." Its 900 pages are characterized by difficult, abstract language and patterns of looping, repetitive phrases ranging from a few words to whole paragraphs. Here is a brief sample of the novel's language:

> Everyone then sometime is a whole one to me, everyone then sometime is a whole one in me, some of these do not for longs times make a whole one to me inside me. Some of them are a whole one in me and then they go to pieces again inside me, repeating comes out of them as pieces to me, pieces of a whole one that only sometimes is a whole one in me.

Not only is *The Making of Americans* little read by the general public, even scholars of Stein's career tend to marginalize it in relation to other shorter and more accessible works. However, the dense yet clearly structured language is ideal for "not-reading" via computational analysis.

Tanya Clement, a graduate student at the University of Maryland who is studying the novel as part of her dissertation, was able to use a variety of digital tools to "not-read" the text. These tools included both popular utilities like "tag clouds" as well as Bradford Paley's beautiful and useful TextArc (www.textarc.org), and various Spotfire™ visualizations. Chief amongst these, however, was a new visualization tool developed at Maryland's Human-Computer Interaction Lab, named FeatureLens [2]. Clement was able to map numerous structures and patterns of repetition throughout the text, and relate them to literary themes, sometimes confirming established interpretations of the novel and sometimes challenging or overturning them. The importance of visualization as a means of accessing and studying the results of the text analysis cannot be over-emphasized. When we look at a painting or picture, we grasp the entirety of it within our optical field. The eye can easily move from one region of the image to the next, looking for patterns and correspondences which aid in interpretation. In the case of a novel (or even a short story or a long poem), however, we cannot hold the entirety of the text within our visual field. Indeed, the physical form of the codex book itself mitigates against this, as the text is arbitrarily broken up into discrete units divided by pages. Visualization, which essentially makes the text a picture, is capable of bringing a novel into focus as a unified visual event. Coupled with rich metadata and the means for computational pattern matching, certain texts like *The Making of Americans*, previously all but unreadable are now *not-readable* in important new ways.

6. Comparison: The MONK Project

The MONK project (Metadata Offers New Knowledge, www.monkproject.org) includes many of the principals from the Nora Project, together with a team from Northwestern which had developed an application called Wordhoard. Wordhoard is a tool for philological study that allows its users to easily juxtapose and compare word usages from widely separated contexts within the same visual field—in effect, using the model of the facing pages of a book, but where each "page" can be arbitrarily placed alongside of any other. Both Nora and Wordhoard depend on comparison, identified by John Unsworth as a scholarly primitive [10], meaning it is a fundamental aspect of what scholars in the humanities do regardless of their particular specialization or object of study. The MONK project collects word-level metadata for every word in its corpus (with plans to scale to a billion words), including lemmatization, part of speech, named entities, and location. It combines this with available metadata related

to a complete work's date and place of publication, the gender of its author, its status as poetry or drama or fiction or non-fiction, and so forth. A typical MONK analytic might consist in seeking to determine the low-level linguistic features of "sentimentality," based on comparative classification and analysis of novels published within the constraints of a particular time period and locale. But MONK itself is envisioned as more of a framework or "plugboard," with any number of applications hanging off of a common data store. Social network analysis is another promising area for literary study which can be implemented using the metadata described above. While MONK is a new project and large portions of its architecture are merely speculative at this time, its design goals clear: to fuse low-level morphological and syntactic metadata for individual words with high-level metadata about the source document. Likewise, MONK is encountering issues and challenges that will be typical of anyone attempting text analysis on this scale, ranging from standardization of the input texts to the limitations of scaling mySQL to designing appropriate UI and visualization conventions, accessible to a non-specialist. What is notable about MONK is its interdisciplinary nature; the participation of humanists as developers and not just as end-users or testers; and (as the project title suggests) its self-consciousness in relation to long-standing traditions of humanistic inquiry.

7. Conclusion

Literary criticism is a non-traditional domain for applications of established techniques in data mining and machine learning. While one cannot interpret the results of a classification of a set of poems in the same way one studies data from a field with known ground truths, the potential to "provoke" a human subject expert may yield insights not readily obtainable otherwise. Furthermore, data mining and associated technologies (like visualization) offer the promise of "not-reading" the vast number of electronic texts that are becoming readily available from a variety of online sources. Far from being a radical departure from previous methods of humanistic inquiry, not-reading and distance reading in fact have their roots in long-standing habits and practices of reading and textual communication. While there will hopefully always be a place for long, leisurely hours spent reading under a tree, this is not the only kind of reading that is meaningful or necessary. Reading is not so much "at risk" as in the process of being remade, both technologically and socially. The digital humanities have important interests as well as expertise in this phenomenon.

8. Acknowledgements

Support for the Nora and MONK research provided by the Andrew W. Mellon foundation. Both projects involved and involve large interdisciplinary teams from universities in the US and Canada; the work described here is the outcome of this collective effort.

9. References

[1] Crane, G. "What Do You Do With a Million Books?" *D-Lib Magazine* 12.3, March 2006. http://www.dlib.org/dlib/march06/crane/03crane.html.

[2] Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., Plaisant, C. "Discovering interesting usage patterns in text collections: integrating text mining with visualization." HCIL Technical Report, 2007-08. http://cgis.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2007-08.

[3] Fischer, S. R. *A History of Reading*. Reaktion Books, London, 2003.

[4] Moretti, F. Graphs, Maps, Trees: *Abstract Models for Literary Theory*. Verso, London, 2005.

[5] Mueller, M. "Notes towards a user manual of MONK." https://apps.lis.uiuc.edu/wiki/display/MONK/Notes+towards+a+user+manual+of+Monk, 2007.

[6] Plaisant, C. and Rose, J. and Yu, B. and Auvil, L. and Kirschenbaum, M. and Smith, M. and Clement, T. and Lord, G., Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces, in *Proc. of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 141-150. 2006.

[7] Ramsay, S. (2004) In Praise of Pattern. In *"The Face of Text" - 3rd Conference of the Canadian Symposium on Text Analysis (CaSTA)*.

[8] Schreibman, S., Siemens, R., and Unsworth, J. (eds.) *A Companion to Digital Humanities*. Blackwell, Oxford, 2004.

[9] Smith, M. N. Email of 11/10/05, subject "Curmudgeon Reflections on nora" to webviz[at]lists.prairienet.org, the project email list for the NORA project.

[10] Unsworth, J., Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? Presented at the *Humanities Computing: formal methods, experimental practice symposium*, King's College, London, May 13, 2000. http://www3.isrl.uiuc.edu/~unsworth/Kings.500/primitives.html.

[11] Zaid, G. *So Many Books: Reading and Publishing in an Age of Abundance*. Paul Dry Books, Philadelphia, 2003.

The â€œreadingâ€ is a form of data mining that allows information in the text or about the text to be processed and analyzed. Debates about distant reading range from the suggestion that it is a misnomer to call it reading, since it is really statistical processing and/or data mining, to arguments that the reading of the corpus of literary or historical (or other) works has a role to play in the humanities.Â Images have different properties in digital form than texts, and the act of remediating an image into a digital file is more radical than the act of typing or transcribing a text into an alphanumeric stream (we could quibble over this, but essentially, text is produced in alphanumeric code, but no equivalent or analogous code exists for images). A new journal, the Journal of Data Mining and Digital Humanities (JDMDH), recently published its inaugural issue. JDMDH is an open access peer-reviewed quarterly journal â€œconcerned with the intersection of computing and the disciplines of the humanities, with tools provided by computing such as data visualisation, information retrieval, statistics, text mining by publishing scholarly work beyond the traditional humanities.â€ The journal is a joint project of CNRS (French National Centre for Scientific Research), INRA (French National Institute for Agricultural Research), and INRIA (French Insti