Scholarship in the Age of Abundance: Enhancing Historical Research With Text-Mining and Analysis Tools

Second Year Report to the National Endowment for the Humanities
August 2010

Dan Cohen & Sean Takats
Center for History and New Media
George Mason University

We have had another busy year on this project as it built upon not only the first year's foundational work in studying the practice of professional historians but also extended into directly related text-mining projects that benefitted strongly from what we have already done. We spent much of the past year creating tools that answer the needs and concerns expressed in the extensive studies that took place in Year 1, and presenting the initial results of our work on this project.

The Smaller Focus Group and the Identification of Two Models

As originally planned, this year we engaged a smaller focus group of historians (a representative sample drawn from the over 200 survey participants from Year 1) in a set of tests. Using small honoria incentives, we established a group of two dozen historians sufficiently diverse in terms of research interests, career stage, and knowledge of technology. The project team gave this group a series of much more involved case studies in the fall of 2009, asking for qualitative responses to different use cases, technologies, methods, and client and server-side software.

After analyzing more closely the commentary from this group, we came to see that they had in mind two major—and disparate—methodologies in their work that could be addressed by text mining: analyzing "slices" of individual corpora that meet certain criteria (e.g., all letters from women in a particular decade), and analyzing "ad hoc" research libraries assembled from multiple online collections (e.g., selected love letters from a variety of archives). As we watched the focus group use certain software, such as SEASR and Voyeur, we tried to indentify the best way (or ways) in which scholarly text-mining programs could to address these two modes.

The first method generally involves a historian asking for all of the documents from a corpus with particular metadata or full-text keyword matches; the second uses a personal assemblage of individual documents, generally on a common topic, drawn from a wide range of collections. In Zotero, for instance, the former is equivalent to a "saved search" on a collection (i.e., all matching items from a Boolean search of a collection); the latter, the folders that a Zotero user can set up in the interface to organize their library into subsections manually. Other tools have similar categories.

We thus began to look at easier ways to pass both slices and ad hoc text corpora to analytical and visualization services to make text mining simpler for the average

historian. For slices, we have come to believe (from tests, e.g., in concert with SEASR) that it is impossible to pull down the entirety of most corpora to a client and then upload it to a service. Instead, it is better to pass a token, URI, or query string to *reference* the text the analytical service needs, which will be then be delivered directly from the collection to the text-mining service. For ad hoc collections, on the other hand, client tools like Zotero can bundle the text itself and send it from the client—if they structure their code properly.

The Creation of New Tools

With these important conclusions in mind from our surveys, we set about creating some new text mining tools that would be easier to use and that understand the functioning of the digital scholarly "triangle" (collection, tool, analytical service). Since both the initial large survey and the focus group seemed particularly attracted to mapping tools, i.e., that extract and map entities from texts, we started there. (Unlike literary text-mining tools, i.e., with pure text or numerical outputs, our historian sample particularly related to spatial elements.) So as a test case to address client-side ad hoc collection mining, we strove to create a mapping tool for Zotero.

The result of this effort, named (uncreatively) Zotero Maps, was released in the spring of 2010 and is now available at http://zotero.org/download/plugins/zoteromaps-1.0.9.xpi (it has been updated numerous times since as the version number suggests). Zotero Maps allows users to select any set of texts in the Zotero client and then uses text mining to extract all entities (e.g., place names) and map them using OpenStreetMap (an open source competitor to Google Maps). Use cases identified in Year 1 of this grant, such as the need to quickly see where a historical subject visited in an assemblage of travel letters, were thus addressed with this tool.

We then took some of the underlying code of Zotero Maps that bundles ad hoc collections of texts to create an alpha of a general tool that can send these texts to any open service that accepts zipped text packages, common-delimited URLs for texts, or free-form text blocks (of which we identified over two dozen in Year 1). We plan to add many text-mining service API URLs to this plugin (yet to be named), so that a simple drop-down menu allows a historian to select what kind of analysis they want, pick their texts, and send with one click. The main web window will be used for results of the text analysis, as with Zotero Maps.

Extending the Work into Other Grants

The other major use case—the "slice"—is being dealt with in work that has extended from this grant into another NEH grant, "Using Zotero and TAPoR on the Old Bailey Proceedings: Data Mining With Criminal Intent," a recipient of a Digging into Data award (the project team thanks the NEH again for this). That project is using some of the server-side techniques we said we would explore in "Scholarship in the Age of Abundance." The Old Bailey Proceedings reflect evidence given in almost two hundred thousand trials at London's central criminal court over a period of 240 years. They comprise 120 million words of structured text, the largest coherent body of printed descriptions of behavior ever published in any form. Despite a wealth of scholarship,

however, historians have tended to approach this collection by sifting through the documents one at a time, using unusual or compelling stories to illustrate social, cultural or intellectual histories. Along with the scholars behind the Canadian-based text analysis services TAPoR and Voyeur, we intend the "Criminal Intent" project to demonstrate potential roles for text mining in historical practice, showing that greater historical rigor can be achieved, and new insights gained, by moving from a single trial or narrow run of relevant examples to an analysis of statistically significant textual patterns found in this source as a single, massive whole.

This grant will leverage the work we have done for "Scholarship in the Age of Abundance" by showing how an end user tool—i.e., the tool most likely to be in the hands of the average working historian, who will not know how to program on a high-performance computer—can pass information *about* a slice of a library or archive website to an analytical service. From our work on "Scholarship," we have developed a technique (as noted above) to pass to the text mining service a simple URI—essentially, something that looks like a URL but actually requests from a digital collection all the documents that match certain criteria. This technique is emerging in commercial services that have APIs, where RESTful services now often employ simple-looking URIs. The Zotero Server has also begun to implement this technique; e.g., http://zotero.org/collection_name will be able to be used to send all the documents in that collection to an analytical service, directly from the Zotero Server to a tool such as Voyeur. We believe, as hoped for in the "Scholarship" proposal, that this technique is extensible to other collections with relatively minimal setup, one of the goals of the grant. We plan to publish specifications for this transfer—using a third party tool only as the passer of a URI token—as part of the conclusion to both "Scholarship" and "Criminal Intent."

We were also fortunate to receive a Google Digital Humanities Research Grant for a project entitled "Reframing the Victorians," which will provide significant help in completing work on the case study on Victorian intellectual history for the "Scholarship" NEH grant. The Google grant will provide access to the tools we need to do the kind of major longitudinal study we had originally envisioned but which we found was difficult with existing tools. Luckily, the vast digital library of Google Books presents for the first time the possibility that we can conduct a comprehensive survey of Victorian writing. We plan to report on the extension of techniques from the NEH grant to the Google Books text-mining system, which we believe will be helpful to other historians considering this work. We also plan to execute this project in a way that will make its methodology, workflow, and toolset extensible to other projects that might revisit long-standing scholarly interpretations or assumptions that can at last be thoroughly tested given corpora such as Google Books.

Dissemination

Project co-lead Sean Takats presented some of his preliminary findings on the other case study from the NEH project plan (on the *Encyclopedie*) at the 2010 meeting of the American Society of Eighteenth-Century Studies in Albuquerque, NM, in a talk entitled "Diderot as Digital Humanist." Takats also presented some of the findings from the overall project at the University of Iowa at a conference on "New Practices of Historical Scholarship" in the fall of 2009. Project co-lead Dan Cohen presented the findings of this

project at the American Historical Association annual meeting in January 2010 in San Diego. He also led a practicum session on text-mining techniques that was well-attended, and also met individually with some of the members of the focus group for additional discussions. Project members Fred Gibbs and Trevor Owens are finishing their article based on the survey and focus group and plan to submit it for peer-reviewed publication this fall.

In the spring we filed for the no-cost extension to complete the work of this grant. This will help us finish some of the proposed writing, release the generalized code for ad hoc collection text mining, and execute remaining deliverables in concert with the Digging into Data and Google grants. We thank the National Endowment for the Humanities again for their generous support of our work.

Sincerely,

Dan Cohen
Sean Takats
Center for History and New Media
George Mason University