THE NORA PROJECT: TEXT MINING AND LITERARY INTERPRETATION

Matthew KIRSCHENBAUM
Panel ABSTRACT

This panel brings together three papers showcasing different facets of the nora Project, a multi-institutional, multi-disciplinary Mellon-funded initiative to apply text mining and visualization techniques to digital humanities text collections.

We are currently one year into the initial two-year phase of the project. Though most of our methods remain tentative, most our findings speculative, and our technical environment experimental, we nonetheless have significant progress to report. In practical terms, work on the project has advanced considerably since the initial demos and research agendas that were presented at last year’s conference (2005). We have conducted four sustained text mining investigations (two of which are discussed in detail in the papers below), built a complete technical environment that allows a non-specialist user to engage in the text mining process, and we have begun to achieve some consistency in our understanding of what data mining in the humanities, particularly literary interpretation, might be good for. While our findings in this last area remain contingent in the extreme, they nonetheless tend to cluster around activities such as provocation, patterning, anomaly, and re-vision (in the most literal sense). In both of the literary test cases documented in the papers in this session, text mining has produced compelling insights that already provide the basis for more traditional scholarly interventions—papers and articles—in their respective subject fields. The technical environments featured in the papers likewise have promise in their own right and stand ready to support text analysis (Tamarind), structured text visualization (Maryland’s adaptation of the InfoVis Toolkit), and a newly designed visual environment in support of the kind of complex, aggregative operations endemic to data mining (the Clear Browser).

All authors listed in the papers have communicated their willingness to participate.

References


“Undiscovered Public Knowledge”: Mining for Patterns of Erotic Language in Emily Dickinson’s Correspondence with Susan Huntington (Gilbert) Dickinson

Catherine PLAISANT
Human Computer Interaction Lab
Martha Nell SMITH
English and MITH, University of Maryland
This paper develops a rationale for “provocational” text mining in literary interpretation; discusses a specific application of the text mining techniques to a corpus of some 200 XML-encoded documents; analyzes the results from the vantage point of a literary scholar with subject expertise; and finally introduces a tool that lets non-specialist users rank a sample set, submit it to a data mining engine, view the results of the classification task, and visualize the interactions of associated metadata using scatterplots and other standard representations.

Text mining, or machine learning as it is also known, is a rapidly expanding field. Canonical applications are classification and clustering (Weiss 2005, Widdows 2004, Witten 2000). These applications are becoming common in industry, as well as defense and law enforcement. They are also increasingly used in the sciences and social sciences, where researchers frequently have very large volumes of data. The humanities, however, are still only just beginning to explore the use of such tools. In the context of the Nora Project, a multidisciplinary team is collaborating to develop an architecture for non-specialists to employ text mining on some 5 GB of 18th and 19th century British and American literature. Just as importantly, however, we are actively working to discover what unique potential these tools might have for the humanist.

While there are undoubtedly opportunities for all of the normative text mining applications in large humanities repositories and digital library collections, their straightforward implementation is not our primary objective with Nora. As Jerome McGann and others have argued, computational methods, in order to make significant inroads into traditional humanities research, must concern themselves directly with matters of interpretation (2001). Our guiding assumption, therefore, has been that our work should be provocational in spirit—rather than vocational, or merely utilitarian—and that the intervention and engagement of a human subject expert is not just a necessary concession to the limits of machine learning but instead an integral part of the interpretative loop. In important respects we see this work as an extension of insights about modeling (McCarty 2004), deformation (McGann 2001), aesthetic provocation (Drucker 2004), and failure (Unsworth 1997). It also comports with some of the earliest applications of data mining, such as when Don Swanson associated magnesium deficiency with migraine headaches, an insight provoked by patterns uncovered by data mining but only subsequently confirmed through a great deal of more traditional medical testing (Heast 1999).

We began with a corpus of about 200 XML-encoded letters comprising correspondence between the poet Emily Dickinson and Susan Huntington (Gilbert) Dickinson, her sister-in-law (married to her brother William Austin). Because debates about what counts as and constitutes the erotic in Dickinson have been primary to study of her work for the last half century, we chose to explore patterns of erotic language in this collection. In a first step our domain expert classified by hand all the documents into two categories “hot” and “not hot.” This was done in order to have a baseline for evaluation of the automatic classifications to be performed later.

We then developed an exploratory prototype tool to allow users to explore automatic classification of documents based on a training set of documents classified manually. The prototype allows users to read a letter and classify it as “hot” or “not-hot” (Fig 1). After manually classifying a representative set of examples (e.g. 15 hot and 15 not-hot documents) this training set is submitted to the data mining classifier. For every other letter in the corpus, users can then see the proposed classification, review the document, and accept or change the proposed classification. The words identified by the data mining as possible indicators of erotic language are highlighted in the text of the document.

Importantly, this process can be performed in an iterative fashion as users improve the training set progressively and re-submit the automatic classification. Currently results are presented in the form of a scatterplot which allows users to see if there is any correlation between the classification and any other metadata attribute of the letters (e.g. date, location, presence of mutilation on the physical document, etc.) Users can see which documents have been classified by hand (they are marked with triangles) and which have been categorized automatically (they appear as a circle). Letters that have been classified as not-hot always appear in black, and in color for hot,
making it easy to rapidly spot the letters of interest.

A key aspect of our work has been to test the feasibility of this fairly complex distributed process. The Web user interface for manual and automatic classification is a Java Web Start application developed at the University of Maryland, based on the InfoVis Toolkit by Jean-Daniel Fekete (2004). It can be launched from a normal Web page and runs on the user’s computer. The automatic classification is performed using a standard Bayesian algorithm executed by a data mining tool called D2K, hosted at the University of Illinois National Center for Supercomputing Applications. A set of web services perform the communication functions between the Java Interface and D2K. The data mining is performed by accessing a Tamarind data store provided by the University of Georgia, which has preprocessed and tokenized the original XML documents. The entire system is now functional.

What of the results? The textual critic Harold Love has observed of “undiscovered public knowledge” (consciously employing the aforementioned Don Swanson’s phrase) that too often knowledge, or its elements, lies (all puns intended) like scattered pieces of a puzzle but remains unknown because its logically related parts are diffused, relationships and correlations suppressed (1993). The word “mine” as a new indicator identified by D2K is exemplary in this regard. Besides possessiveness, “mine” connotes delving deep, plumbing, penetrating—things we associate with the erotic at one point or another. So “mine” should have already been identified as a “likely hot” word, but has not been, oddly enough, in the extensive critical literature on Dickinson’s desires. “Vinnie” (Dickinson’s sister Lavinia) was also labeled by the data mining classifier as one of the top five “hot” words. At first, this word appeared to be a mistake, a choice based on proximity to words that are actually erotic. 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 full of the daily details of life in the Dickinson household. While extensive writing has been done on the blending of the erotic with the domestic, of the familial with the erotic, and so forth, the determination that “Vinnie” in and of itself was just as erotic as words like “mine” or “write” was illuminating. The result was a reminder of how or why some words are considered erotic: by their relationship to other words. While a scholar may un-self-consciously divide epistolary subjects within the same letter, sometimes within a sentence or two of one another, into completely separate categories, the data mining classifier will not. Remembering Dickinson’s “A pen has so many inflections and a voice but one,” the data mining has made us, in the words of our subject expert, “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.”
Figure 3: This is an alternate "scatterplot" view of the collection. Each dot represents a document. Time (i.e., the median of the estimated date range) is mapped on the X axis, and the length of the document is mapped on the Y axis. Color represents hotness. The same color coding is used. We can see that the longer documents were written earlier on. The display also suggests that there is no correlation between time and hotness, and no particular time periods where significantly more hot documents were written. Zooming is possible to inspect particular clusters.

References


Distinguished Speakers: Keyword Extraction and Critical Analysis with Virginia Woolf’s The Waves

Stephen RAMSAY
Sara STEGER
Department of English, University of Georgia

From the earliest epistolary novels of the eighteenth century to the stream-of-consciousness narratives of the twentieth, English novelists have constructed narratives in which a single story is told from a variety of different first-person viewpoints. The motivations...
for this technique are as ramified as the variations of the technique itself; some authors have used it to demonstrate the contingent nature of subjectivity, while others have employed the technique merely as a way of increasing dramatic irony and tension. In most cases, the individuated chorus of speakers is distinguished stylistically. One might compare, in this context, the scientific formality of the male characters in Dracula with the more personable journalistic endeavors of Mina and the twitting effusions of Lucy; the unprepossessing nobility of Hartright with the urbane wickedness of the Count in The Woman in White; the fractured, acausal narrative of Benjy Compson with the neurotic eloquence of his brother Quentin in The Sound and the Fury.

Virginia Woolf’s 1931 novel The Waves employs this approach to narrative in order to trace the lives of six friends from early childhood to old age. The characters each tell their stories at seven distinct stages of their lives. Each monologue is clearly delineated, and all six characters share some of the same experiences at different points in the narrative.

Yet some critics claim these characters are not differentiated from one another stylistically, and that the distinguishing features of the characters have more to do with the complex symbolic landscape each one inhabits -- a motif foregrounded by the lack of stylistic differentiation. J. Guiget, for example, maintains that “these are not voices, in the sense that they are not differentiated. But for the ‘Bernard said’ or ‘Jinny said’ that introduces them, they would be indistinguishable; they have the same texture, the same substance, the same tone” (283). Likewise M. Rosenthal suggests that although the speeches are highly stylized, the language is undifferentiated. He goes on to say that “although the different speakers all have different points of view and preoccupations, they use the same kind of sentence rhythms and employ similar kinds of image patterns” (144). For these critics, the six characters are not characters at all, but voices indistinguishable by means of language or imagery.

There is, however, an opposing critical position that stresses the importance of stylistic differentiation among the characters in the novel (the specific contours of this difference being the main point of scholarly contention). Charlotte Mendez draws the line of differentiation along the gender axis. Hermione Lee emphasizes the distinction between the worldly speakers and the more solitary speakers. Susan Gorsky emphasizes the individuality of the characters through a clustering of their primary images. According to her research, “the speech of each character is made distinct within the mask of the formal monologue by the repetition of key phrases and images. Diction varies from one speaker to the next because of the words repeated in the image patterns” (454).

With this critical backdrop, certain questions naturally emerge. Do the characters employ similar image patterns or distinctive language patterns? Is there a way to group characters based on similarities in their speeches? Are there six voices in the novel or is there only one?

Our goal was not to adjudicate the matter, but to seek further entry points into these axes of intelligibility. Knowing that vocabulary (symbolic or not) would be one vector of interest, we employed several variations of the tf-idf formula as a way to separate the characters. Tf-idf -- a popular formula in information retrieval -- weights term frequency in a document against the frequency of that term throughout a corpus. By assigning weights to terms, it attempts to re-fit a word frequency list so that terms are not distributed according to Zipf’s law. We compared every word token in each of the six characters to every other character’s vocabulary, and used the resulting lists of “distinctive terms” as the basis for further reflection on the individuation of character in Woolf’s novel.

In generating our results, we had recourse to Tamarind (one of the software subsystems for the nora project). This system, which acts as an XML pre-processor for scholarly text analysis, tokenizes, parses, and determines part-of-speech markers for each distinct token in a corpus. Using this system as a base, we were able to conduct comprehensive term-comparisons using only 50 lines of code (in Common Lisp). We therefore think of this project as a test case for the feasibility of using Tamarind as way to simplify complex text analysis procedures of the sort envisioned by the larger nora project.

In this paper, we present the results of our investigation into Woolf’s narrative, while also looking at the ways in which the software architecture for nora enabled us to undertake the study quickly and easily. Drawing on similar work with tf-idf in digital humanities (e.g. Rydberg-Cox’s work with Ancient Greek literature), we suggest some of the ways in which the results of keyword
extraction algorithms might be further processed and visualized. Finally, and perhaps most importantly, we discuss the ways in which the computerized generation of “suggestive pattern” can enable critical reflection in literary study.

References


The Clear Browser: Visually Positioning an Interface for Data Mining by Humanities Scholars

Stan RUECKER
Humanities Computing in English and Film Studies, University

Ximena ROSSELLO
Dept of Art and Design, University of Alberta

Greg LORD
Maryland Institute for Technology in the Humanities (MITH), University of Maryland

Milena RADZIKOWSKA
Centre for Communication Studies, Mount Royal College

We describe in this paper a strategy for interface design based on the concept of visual positioning. We apply this strategy to the design of an interface for the Nora project, which presents a unique opportunity to create tools to accommodate a powerful technology-data mining-to a new group of users-humanities scholars.

The goal of the Nora project is to apply state-of-the-art data mining processes to a wide range of problems in the humanities (Unsworth 2005), not only in the service of hypothesis testing, but also as a means of contributing to hypothesis formulation (Shneiderman 2001; Ramsay 2003). In both of these cases, however, the question arises of how to make the power of data mining for text collections accessible to academics who are neither mathematicians nor computer programmers. Typical interfaces for data mining operations involve either command lines, such as are used in working in UNIX, or else GUIs, the visual positioning of which frequently places them in a technical domain-many resembling the interfaces used in software compilers. For humanities scholars, it is necessary to consider alternative designs that attempt to adopt a visual position that is at once more congenial and more appropriate for humanists, while at the same time sacrificing as little as possible of the functional control of the underlying system.

The concept of visual positioning has become widespread in the visual communication design community. An early formulation of the principle was provided by Frascara (1997) who pointed out that since one of the primary goals of the graphic designer is to improve communication, it is necessary to consider the visual environment and visual preferences of the users in order to increase the success of the design in communicating with them. The application of this concept to interface design suggests that there are going to be designs that are more or less successful for a particular group of users, and that the same designs won’t necessarily be successful to the same degree with a different group that does not share the same visual position.

In connection with the Nora project, the necessary communication is between the technical mechanism
of the data mining processes and the potential user-the humanities scholar. A typical data mining operation consists of the following stages:

1) the system provides the user (in this case, a scholar) with a sample of documents from the collection

2) the scholar chooses among the sample documents those which are of interest for a particular study. In the two Nora project examples, a sample of poems from a collection of Emily Dickinson was rated in terms of erotic content, and a sample of novel chapters was rated according to their instantiation of the concept “sentimentalism.”

3) the system performs a set of “feature extraction” actions in order to determine shared characteristics of the selected documents

4) the scholar examines the shared characteristics and iteratively adjusts the result as necessary

5) the system applies the resolved characteristics to the larger collection in order to automatically identify similar documents

6) the scholar studies both the shared characteristics and the result set, often by using a visualization tool (in Nora, the InfoVis toolkit).

We call the interface intended to facilitate this process the clear browser. It is based on the idea of rich-prospect browsing, where some meaningful representation of every item in the collection is combined with a set of tools for manipulating the display (Ruecker 2003). In this case, the primary tools are in the form of a set of “kernels” which encapsulate in visual form the results of the data training stage. The kernels allow a simple means of storing the results of feature extraction processes for further modification or use, and also give the user a simple mechanism for applying the process, by dragging and dropping the kernel within the representation of all the collection items (Figure 1). The effects of the kernel are to visually subset the collection items into two groups—selected and unselected—so that the user can subsequently access the items in the selected subset. The design also allows for combinations of kernels, and for a single kernel to provide multiple functions, including not only subsetting the items, but also adding further grouping or sorting functions, as well as changes to the form of representation.

One of the important aspects of the visual positioning for humanities scholars is the proposed form of the meaningful representation of the individual items in the collection. These items are each a piece of text, and together they form a large body of text that is displayed on screen as the default interface. It perhaps goes without saying that humanities scholars are comfortable with text, whether in print or on screens, and the choice to represent collection items with text can therefore contribute to their ability to interpret quickly and intuitively what is happening with a system that might otherwise be unfamiliar or disorienting.

For purposes of illustration, it might be helpful at this point to introduce a scenario involving changes to the form of representation. Such a change might be introduced by the system in connection with a sorting action. For example, if the items in the collection are initially represented as the titles of poems, and the user elects to sort the selected poems by date of first publication, it would typically be useful at that point to add the date to the name of each poem. This addition would constitute a change to the individual representations of items. Alternatively, in cases where the user prefers to group the items rather than sort them,
the additional information might be attached to the entire group in the form of a group label, in which case the representations of the individual items in the group would remain unchanged.

Another aspect of the visual positioning is the animated actions of the kernels, which interact with the field of representations with an effect like oil and water. The animation of the movement of the text items, which move to the periphery of the display or the centre of the area associated with the kernel, provides two kinds of cognitive reassurance. First, the user has a sense of being able to follow the action of the data mining process as encapsulated in the kernel. Second, the animated transitions of the text items provide reassurance that the system is rearranging the collection without adding or subtracting any items. This second factor is particularly important in cases where one of the other functions of the kernel is to add or subtract components from the meaningful representation. By animating the movements and changes in discrete steps, the interface helps make the results of the process understandable. The animated actions of the items become part of the visual positioning, not because cognitive reassurance isn’t important for all users, but because some users can benefit more than others from having it provided in this form.

References


Comparing Aggregate Syntaxes

John NERBONNE
University of Groningen, The Netherlands

Franz MANNI
Musée de l’Homme, Paris

Recently, large and representative databases have become available which record how languages and dialects differ syntactically, i.e. with respect to the way in which words and phrases combine. Barbiers, Cornips and van der Kleij (2002) and Longobardi are examples of such data collections, and several more are in construction, and should be available presently. This availability of large amounts of digitized, controlled syntactic data enables several new questions to be addressed, including how syntactic features are geographically distributed (dialectology and geolinguistics, see Spruit, 2005), to what degree syntactic features associate with one another (typology), and the degree to which syntactic features follow phonological and lexical features in their geographic or linguistic (typological) patterning.

Perhaps most intriguingly both Guardino & Longobardi (2005) and Dunn et al. (2005) postulate that syntactic features are more resistant to change than other linguistic properties, so that careful examination of shared syntactic features and especially shared syntactic innovation should be not only a welcome addition to the techniques of historical reconstruction in linguistics, which rely primarily on phonetic and morphological evidence, but potentially an improvement.

None of these investigations is conceivable without extensive computational support. Not only is the data digitized, but the analytical procedures, the statistical analysis of the results and their visualization all require significant processing time.

The purpose of this special session will be to present representative studies exploring the various ways in which these syntactic databases are being exploited, including ways which might engage other students of the history of human culture.

There will be three papers, one each by the groups at