NORA annual report for the period October 2004 through October 2005

There are six parts to this report:

1. Participants
2. Purpose
3. Materials
4. Architecture
5. Accomplishments to Date and Future Plans, by Participating Institutions
6. Reference URLs

1. Participants:

There are no participants who are funded full-time for their work in this project. The participants funded for part of their time are at five different institutions. Those participants are: Stephen Ramsay and Sara Steger, University of Georgia; Stan Ruecker, Ximena Rossello, Milena Radzikowska, and Bernie Roessler, University of Alberta, Tom Horton, Worthy Martin, Kristen Taylor, Ben Taitelbaum, Charles Zhang, University of Virginia; Tanya Clement, Matt Kirschenbaum, Greg Lord, Catherine Plaisant, Martha Nell Smith, James Rose, University of Maryland; Loretta Auvil, John Unsworth, Bei Yu, Xin Xiang, University of Illinois. The disciplinary expertise represented across this group includes literary scholarship, data-mining/machine learning, software engineering, graphic design, human-computer interaction, and library & information science.

2. Purpose

In spite of the new technologies with which it works, the purpose of the NORA project is "old school," as Matt Kirschenbaum has put it: it aims to facilitate one of the oldest activities in the humanities, and indeed one of the earliest cognitive abilities of humans, namely pattern recognition. Specifically, we aim at allowing literary scholars to identify categories of interest in collections of literary texts, and then to find new members of those categories and explore the features that correlate with those categories—whether across the works of a particular author, era, genre, or simply across the collection as a whole. This will represent a significant advance beyond the familiar keyword searching as a mode of access to digital libraries, because it will allow researchers to do things like query by example ("find me more texts like this one") or to explore questions like "what makes a novel sentimental?" The purpose of doing this is not so much to resolve questions but to raise them—in other words, we expect that if we are successful, the results of the process may well be published as a traditional scholarly article or essay (albeit with some illustrations of an unusual sort) in which an argument is carried out, based on evidence *discovered* by using the NORA tools.

3. Materials:

We began by collecting XML- and SGML-encoded British and American 18th- and 19th-century literary texts, from libraries (at The University of North Carolina-Chapel Hill, the University of Virginia, the University of Michigan, Indiana University, University of California at Davis, and the Library of Congress) and from scholarly projects (The Perseus Project, The Brown University Women Writers Project, The Walt Whitman Archive, the Emily Dickinson Electronic Archives, the Rossetti Archive, the Blake Archive, Uncle Tom's Cabin and American Culture). Altogether, we have collected about 5 GB of marked-up machine-readable literary text—about 10,000 individual texts. Almost all texts have been encoded according to the Text Encoding Initiative Guidelines, but there is a great deal of room for variation in encoding practice within the TEI Guidelines, and these texts exhibit considerable variation both within and across collections. The major difference is between library texts and the texts from scholarly projects: the latter are much more heavily encoded and their encoding reflects specific interests or purposes with respect to the texts. Neither the texts from scholarly projects nor the texts from libraries were prepared with the expectation that they would be processed or used outside the original system context in which they were designed to be published, so they lack some useful information from our point of view—for example, most files don't include, within the file, their own filename, or a formal public identifier that could be resolved to the published version of the file.

4. Architecture

Software architecture has been one of the major challenges of the NORA project, so far. We still think it is a reasonable expectation that we will be able to produce working tools for text-mining within the two-year period of this grant, and that is in part because of the head start we get by using D2K for the basic data-mining processes, but as Figure 1 shows, there are many other parts to NORA beyond D2K. The basic components are

- D2K (developed and still developing at NCSA),
- Tamarind (our relational database management system, being built for this project by Steve Ramsay and Sara Steger at the University of Georgia),
- the visualization of results, which also needs to function as a user interface (being developed by participants at Maryland, using a Java toolkit for information visualization that comes from the University of Montreal, and with some design assistance from participants at Alberta),
- an index of available texts (which we don't currently use, but which would look like an off-the-shelf xml search engine and index, and which may be built into Tamarind by the completion of the project),
- and what's labeled "Backend" in the diagram above, which is where queries, intermediate data sets, state of the visualization/interface, and other kinds of temporary information generated in the process of iterative data exploration is stored (at this point, most of this is actually either canned in advance or stored client-side, in the Java visualization).
- Connecting these elements, in Figure 1, are Web Services (a formal way of specifying interactive services that can be run over the Web).

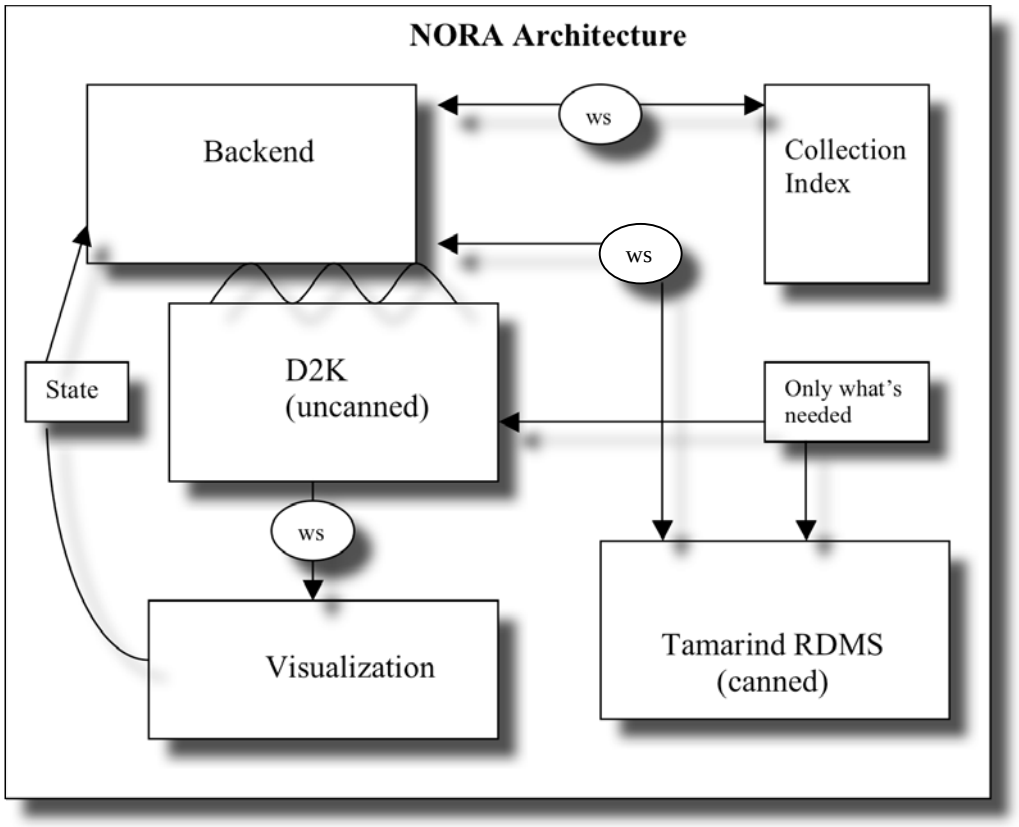

Figure 1: Architecture

The information in Tamarind is "canned" in the sense that Tamarind contains full-text collections that have been preprocessed for part of speech tagging, XPath generation, and other time-consuming processing tasks, which is necessary if we are to deliver interactive access to extremely granular quantitative information about large collections. The information that gets passed from Tamarind to D2K, for the actual data-mining, should be "only what's needed" because this exchange of information is a very real bottleneck in the architecture: a query that asks for the frequency of all tokens in a relatively small collection of documents (like all of Emily Dickinson's poetry) might produce a matrix with millions of cells, many of them empty. Sparse matrix data of this sort is a well understood problem in text-mining, but figuring out how to construct our text-mining scenarios so that we don't pass around lots of data that will never be used, and figuring out how to identify the data that will be used and pass it with maximum efficiency, are problems that we have to solve in the context of our particular uses, collections, and architecture. Even in the supercomputing community represented by NCSA, real-time interactive data-mining is recognized as a bleeding-edge challenge.

Another, more general challenge to the NORA project has been the tension between the desire to develop working software and the need for rapid and iterative experimentation. The problem is that if significant pieces of the project are developed in isolation, without the constraint of interoperating with other pieces, we could easily end up being unable to assemble a working whole. On the other hand, if we have to design and implement all the pieces and assemble them into a working whole before we can begin to experiment with using text-mining on literary collections, we run the risk of developing software that doesn't answer to the real interests of scholars, or that isn't up to the real rigors of collection-level data exploration. Our compromise between these two extremes has been to mock up some early demos without critical parts of the NORA architecture having been developed, and to push for integrating these demos into the developing architecture as soon as possible. As a result, sometimes experimentation with text-

mining problems, visualization strategies, and interface design has been delayed by the need to bring a piece of the architecture online, and development of the architecture has sometimes been complicated by new requirements that emerge out of tinkering with demos. We don't really see any alternative to this sort of compromise, and we do not regard it as a problem with the project: indeed, keeping these forces in tension is probably a good thing.

We are pleased, therefore, to report that at the end of the first year of the project we have an end-to-end working example of the NORA architecture, minus the collection index. In other words, although we select the collection subsets by hand, having made that selection we can now run entire text-mining operations involving user input, requests for quantitative information delivered from the user-interface to Tamarind, with results coming back from Tamarind to D2K, and output being delivered back to the user-interface. You can see this in action at:

http://www.noraproject.org/demos.php

All of this can take place on one machine or on several (as it does in this demo, the user-interface software runs as a Java Webstart application delivered to the end-user's desktop from a server in Maryland, with Tamarind running on the NORA server and D2K running on an NCSA server). We can also run Tamarind and D2K on the same server, and in principle we could run the user interface as an applet on that same server, rather than as an application on the client machine. D2K also allows us to parallelize computing-intensive data-mining processes, if necessary. But most importantly, now that these pieces are actually all hooked up, we can rapidly devise, deploy, and evaluate new text-mining demonstrations and strategies, simply by selecting new document sets and writing new data-mining "itineraries" in D2K's easy-to-use graphical programming environment.

In the coming year, we'll experiment with different process distribution strategies to see what produces the best performance, and we will either integrate an off-the-shelf collection indexing engine (like Apache's Lucene) or build that capability into Tamarind, depending on what seems likely to produce the best results with the least complexity for those who manage collections. We're also working on formalizing the "back end" portion of the architecture, and on building some small but critical components that are still missing, like a module for Tamarind that will export sparse matrix information to D2K, and some optional pieces, like a module for D2K that will encapsulate the clustering techniques developed at the National Institute for Technology in Liberal Education.

5. Accomplishments to Date and Future Plans, by Participating Institutions

In general, it is a mark of maturation and success in the NORA project that the graduate students at each of the five sites are now able to make some of the most important contributions to the project, and regularly work directly with one another, rather than having their participation mediated by faculty participants. This is significant because it indicates that the overall project direction and the overall software environment is now mature enough that it is clear what pieces need to be built or tested or populated with data, and in many cases it is clear how to do that, without high-level decision-making. Major contributions have been made to the project in recent months by Tanya Clements, James Rose, and Greg Lord at Maryland, by Sara Steger at Georgia, by Bei Yu and Xin Xiang at Illinois, by Ximena Rossello at Alberta, and by Kristen Taylor at Virginia. We expect even greater contributions in the final year of the project, from these and other students, particularly since, in many cases, the work they are doing on the NORA project is directly related to their own graduate research and training.

University of Alberta:

Stan Ruecker at the University of Alberta was funded as a participant in the NORA project late in the first year of the project, and he and his students have only been actively working with the rest of the NORA team since the fall of 2005. The purpose of the Alberta sub-project is to complement the visualization research already underway at the University of Maryland, developing specific components of NORA's interface and visualization tools. The particular strength of the group in Alberta is in graphic design and usability in software interfaces, and this complements strengths at Maryland in information visualization and human-computer interaction.

The basic thrust of the work at Alberta follows the notion of rich-prospect browsing, where the default interface contains some meaningful representation of every item in the collection, or of every item in the preliminary search results, combined with tools for manipulating the display. Rich-prospect browsing interfaces are typically a form of Zoomable User Interface. Where possible, they should contain the following features:

1. The default interface displays some meaningful representation of every item in the collection
2. The user can manipulate the display using tools designed to organize or change the representations
3. If possible, the tools should be emergent from the information available in the collection
4. If possible, more than one representation should be available for each item
5. The visualization should be the interface to further data and additional tools

There are several conceptual areas to explore in this part of the NORA project, including the need to provide users not only with a prospect view and tools for a particular collection, but also for documents aggregated from multiple collections. An additional level of complexity is introduced by the presence or absence of specific document metadata and interpretive-level markup, and by the provision through Tamarind of a wide range of data mining formulas that can be applied as a means of sorting, subsetting, and otherwise grouping the collection items.

Several visualization concepts are being prototyped in an iterative cycle that includes preliminary paper sketches, refined digital sketches or storyboards (in Photoshop or Illustrator), and interactive sketches (in Flash). The interactive sketches will then be developed into working user interfaces at Maryland and other NORA sites. For these interfaces, the variety and usefulness of the meaningful representations will be a significant success factor. It should be possible to provide microtexts that contain information such as document authors, titles, keywords, facets, and other relevant forms of representation that can be modified by the user, and expanded or contracted as necessary. The design details of representation are also significant, since manifest attention to detail can help to inspire the confidence of the user.

In the coming year, a NORA browser for showing the results of applying data-mining formula will extend previous work on Photofinder (Shneiderman et al. 2002) and the pill identification system (Ruecker et al. 2005), which allows users to look for pill information by browsing a display of pill photos. The working title for this project is the Clear Browser. It will provide a visual field consisting of representations of collection items that are visually re-organized by dragging repelling kernels into the main window. Repelling kernels serve to help clear the display of unwanted items by pushing them out toward the periphery. Each kernel represents a data-mining operation combined with functions to organize the selected items into groups or sorted lists, and to select what kind of information to display for the items.

In this browser, one challenge will be to provide the user with sufficient insight into the formula that is represented by each kernel. Since mathematical equations are not typically meaningful for humanities scholars, other forms of representation will need to be developed. These representations will need to be interactive, so that the parameters of the data-mining operation are under the control of the user.

Over the coming months, the participants at Alberta will put together a set of designs that accommodate the six different kinds of features (words, sentences, s, etc.) that the NORA group has agreed upon as basic units of investigation in literary texts, and they will be looking at interfaces for both batch and interactive processing, since with large collections some batch processing may still be necessary.

University of Georgia:

Because NORA is designed to work on entire document collections, it requires the kind of intensive processing, large-scale storage, and high-volume data transfer associated with corpus-level analysis. But unlike most systems, NORA is designed to perform this kind of analysis online and with a high degree of user configurability. We anticipate some operations facilitated by NORA being no more complex than using a search engine, but in the end, scholars are only going to trust data analysis procedures that they can control themselves.

Giving users a high level of control presents us with some extremely difficult technical challenges. Most software systems can be usefully viewed as separating knowledge about the state of the system between "compile-time" and "runtime." Compile-time knowledge is whatever information you have about the state of the system before the user interacts with it. If you have a lot of information about the state of the data at compile-time, you can optimize the data to ensure that when the user does interact with it, it's processed and delivered in an extremely efficient manner. Runtime data, by contrast, is whatever is known about the state of the system when the program is executing. Programmers will speak of a piece of information being "only known at runtime," which is another way of saying that their ability to optimize the system is limited by the fact that they don't know the size and shape of data (or even what data is in question) until the program is actually run.

Moving corpus-level data mining from an offline activity to an online one means moving an enormous amount of information from compile-time into run-time. In the case of NORA, the system might not know what subset of documents the user wants to work on, what particular parameters the user is interested in, and what particular procedure the user wants to run. For example, in a typical use scenario, the user might select a subset of documents to process, a set of parameters to look at (for example, the number of scenes in play, the ratio of nouns to verbs within paragraphs, the total length of the document, and so forth), and a particular data-mining procedure for analyzing the data. If we were working on the original XML documents, those documents would need to be ingested into the system, parsed, disassembled into their component parts, and processed in such a way as to extract whatever quantitative information corresponded to the parameters of the request. All of this would have to happen, moreover, before the data-mining algorithm (itself an intensive process) had even begun. We are aware of no system—even based in dedicated hardware—that could perform such an operation within the acceptable time-limits of the World Wide Web. Even a modest collection could take days to process, from start to finish, for an arbitrary investigative purpose.

Our solution, therefore, was to develop an extremely aggressive approach to the optimization of any data that might be needed prior to the commencement of the data-mining procedure itself. The result of that approach is Tamarind—one of the three main sub-components of the NORA

architecture. Tamarind's main job is the extraction of useful data from plain XML documents. We began by isolating a core set of features that we think could be generally useful for text mining. Tamarind records all structural information pertaining to the XML markup in the original document, for example. We also decided that individual words would constitute a major area of interest (even though we knew that we would eventually want to operate on larger textual units). For most types of literary and linguistic analysis, the distinctions between word-types represent a basic area of interest, and so we decided that part-of-speech, orthographical classification, and token type (words, punctuation marks, etc.) would also be recorded. Since such information is not usually marked explicitly in a tagged document, we knew we would need to develop a system that could make these determinations separately.

The Tamarind development team at Georgia developed a method for generating XPath expressions for all of the components of ordinary and arbitrary XML documents (XPath is a standard notation for representing individual parts of an XML document), and wrote a tool that can generate these expressions. Rather than develop our own part-of-speech tagger, tokenizer, and type-classification tool, we turned to a mature development framework called GATE (General Architecture for Text Engineering), developed by the Natural Language Processing Group at Sheffield University. With the XPath generator in place, and the GATE libraries re-configured to process our documents, we then proceeded to write a system that could take all of this information and load it into a database schema optimized for efficient retrieval. We then wrote a number of functions within the database system itself that could further analyze the imported data and generate various kinds of quantitative information about the extracted data (for example, frequencies for each extracted feature). We decided to use an ordinary relational database (PostgreSQL, a popular open source tool) as the data storage mechanism, so we could take advantage of the expressive power of SQL (the standard query language for relational database) in formulating methods of search and retrieval.

We experimented with a number of methods for exposing this database to outside clients, including a web-services framework that allows clients (written in virtually any language) to communicate with the datastore over a network connection. We also experimented with more traditional forms of access (like exposing the database as a local library). The web services architecture proved extremely easy to use.

Next, we wrote a number of programs (using a variety of programming language) as demonstrations of Tamarind's utility as a text analysis pre-processor, and each one was able to import data from Tamarind quickly using only a few lines of code. Unfortunately, this method only seems feasible for relatively modest data requests, since HTTP (the standard protocol of the World Wide Web) was not designed for high-speed transfer of large volumes of data. For this reason, it seems likely that in the final system, D2K will access the Tamarind datastore locally. Still, we think that the web services version of Tamarind may be an extremely useful tool for many types of text analysis outside of the context of D2K. Using Tamarind, scholars and students developing or using text analysis software will be able to undertake complex analytical procedures without having to perform some of the more tedious and intensive aspects of the process (since Tamarind has done most of the heavy lifting involved with such procedures already). The main developer of the Tamarind system, Stephen Ramsay, plans to explore this use of Tamarind in his humanities computing classes in Spring of 2006.

Eight months into the project, we had working system that included all the features just mentioned. The actual loading of the datastore was painfully slow (on the order of days for some of our document collections)—even after a month of optimizing the system for efficient use of memory and system resources. But as we predicted, the delivery of user-specified data to the processing layer (D2K) took only a few seconds. Requests on some of the larger datasets in our testbed, however, still tax the system too heavily (moving retrieval times back into the range of several minutes). Thus, one year into the project, we find ourselves with a complete, working

tool, but one that still requires considerable performance analysis and optimization before it will be usable in the real world.

We are not aware of any system that works exactly like Tamarind, and some of its internal structures appear to represent unique approaches the problem of XML storage and retrieval. We would like to continue adding features to the system (particularly features that allow the user to process larger text structures within individual documents), but we are trying to make sure that the demand for new features doesn't override the central purpose of Tamarind—namely keeping the overall NORA system fast, lightweight, and useful for literary text-mining.

In addition to work on Tamarind, participants at Georgia have also presented papers at CaSTA 2004, ACH/ALLC 2005, and MLA 2005, and participated in a panel proposal for Digital Humanities 2006. Ramsay and Steger have also regularly participated in conference calls and all-hands meetings, as well as in regular discussion on the email list, posting documentation to the wiki, with much time spent in instant messaging discussions as well.

University of Maryland:

Maryland's multidisciplinary team includes humanity scholars, a user-interface expert and a computer science graduate student: they meet weekly and their focus has been on designing text-mining experiments and developing interactive visualizations of text-mining processes and results. In the first year of the NORA project, this group began by identifying a specific research topic of interest to scholars at Maryland (the study of the erotics in Emily Dickinson letters), and used this to drive development of two exploratory interface prototypes.

Prototype 1:

Maryland's first exploratory prototype was developed in the spring of 2005 and served as a platform to elicit comments and suggestions from scholars (Fig 2). The prototypes allows users to search for words in a collection and see a display of the locations of those words on an visual overview of the collection, giving users an idea of how many times and where the word occurs in the collection. Users can then click on an occurrence in the overview and see the full text of the letter with the word highlighted. We explored how combinations of words could be entered and how they would be displayed with multiple colors or an overall rating on the text.

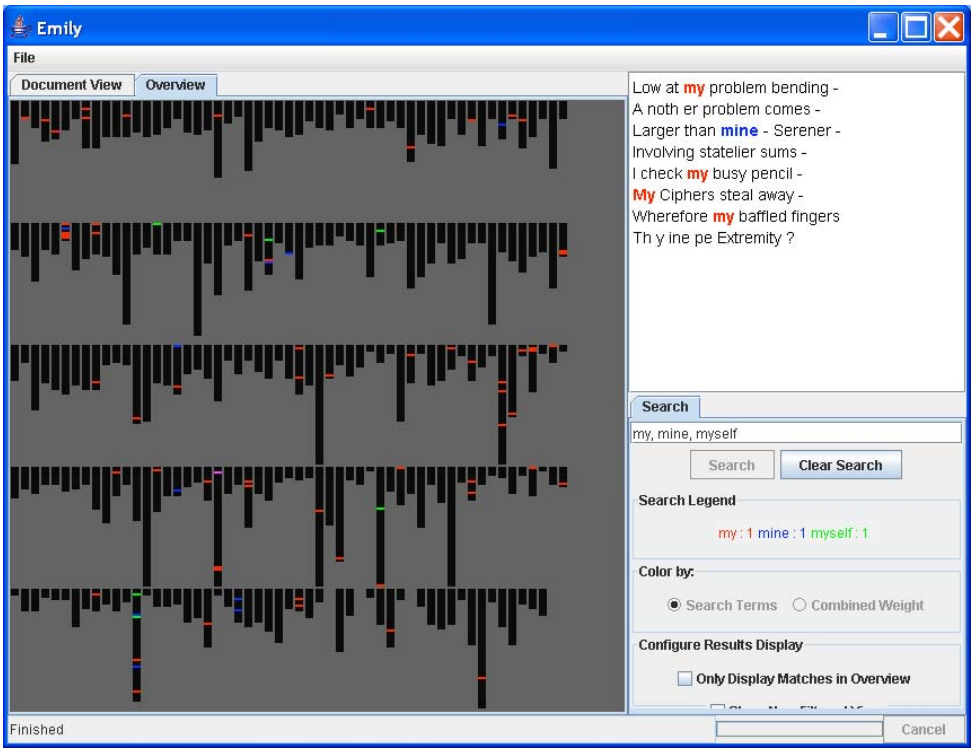

Fig. 2: A first exploratory prototype.

All letters are represented at once on the left. Each bar is a letter, the size of the bar is proportional to the number of lines in the letter (longer letters have longer bars). Clicking on a bar display the text of the letter on the right. Those bars could possibly be ordered by the date of the letter. Here a user has typed a set of 3 words, and the location of the words in the letters is indicated on the bar (giving a approximate location of the word in the letters). Different colors are assigned to different words.

Prototype 2:

We then started to develop a second exploratory prototype to allow users to:

1. *explore document-level metadata to find patterns and correlations between documents descriptors available*
2. *use D2K data mining to explore automatic classification of documents based on a training set classified manually.*

1) Exploration of document-level metadata

The interface takes as input a metadata table requested from D2K. It contains a combination of document descriptors extracted from the XML tags (e.g. date) or computed by Tamarind (e.g. word counts). In this first step a simple visualization is used, namely a scatterplot. We recognize that scatterplots are not intuitively informative for literary scholars: the point in this prototype was simply to demonstrate that metadata from a combination of sources could be delivered into an interactive visualization. The interface allows users to select two attributes and produce an overview of the entire collection where every item is a document of the collection (e.g. a letter in the Emily Dickinson letter) (Figs. 3 and 4)

Fig. 3: Each Emily Dickinson letter is a icon on the scatterplot (a square).
A complete manual classification of hot vs. not hot was done by hand. The hot documents appear a darker red than the not-hot document. Users can choose what attribute is shown on the X and Y axis. Here the X-axis is time (using an XML date tag) and the Y axis is the word count (computed by Tamarind).

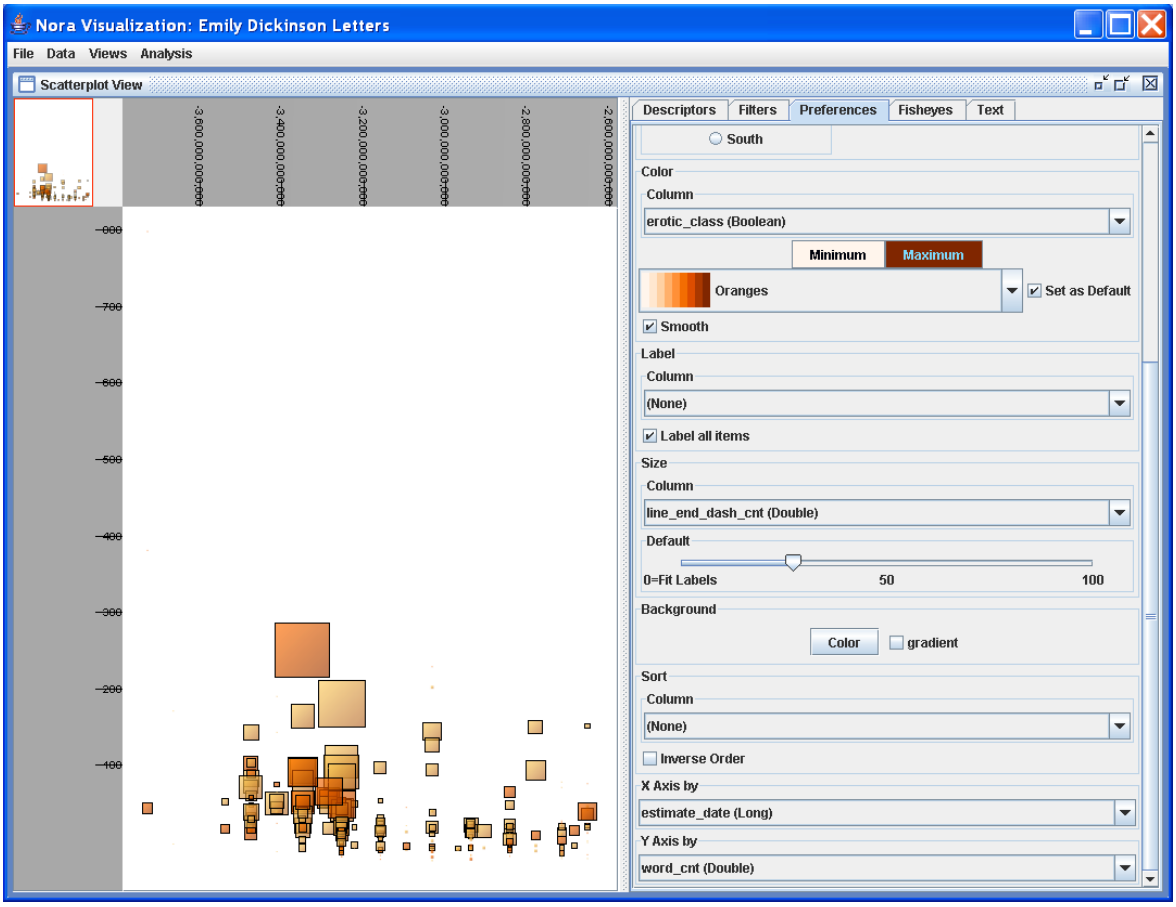

Fig. 4: The size of the square is proportional to the count of end-of-line dashes.

This display suggests that the end-of line dashes were used more often during a certain time period (i.e. many large dots are grouped on a subset of the time axis, the X axis). On the other hand, large dots are neither more dark red than pale, so there does not seem to be an obvious correlation between the dashes and "hotness". One document stands out as an outlier having a lot of those dashes. Clicking on that dot will bring up the text of the poem in the right window.

2) Automatic classification via D2K

The prototype allows users to classify documents manually (e,g, as hot or not-hot) (Fig 5) and submit the set of classified documents as a training set for the D2K data-mining classifier (Fig 6). Finally users can review the suggestions of the D2K classifier (Fig. 7) and accept or reject the results. The next step will be to allow iterative refinement of the classification. We can also provide alternate views of the results. For example, a simple list of documents enhanced with color icons will facilitate the systematic review of the results. The multiplicity of views allows users to see the results in different ways.

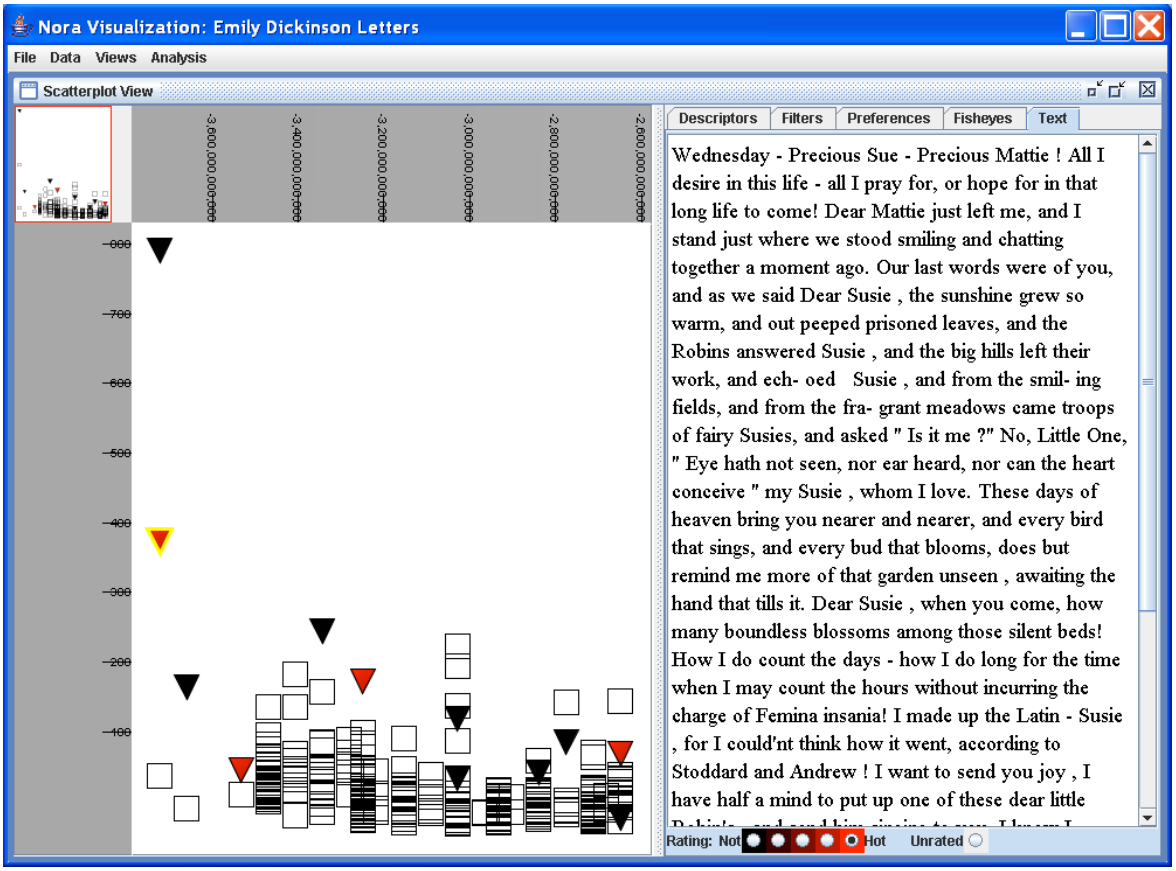

Fig. 5: Manual classification of a training set of Dickinson letters

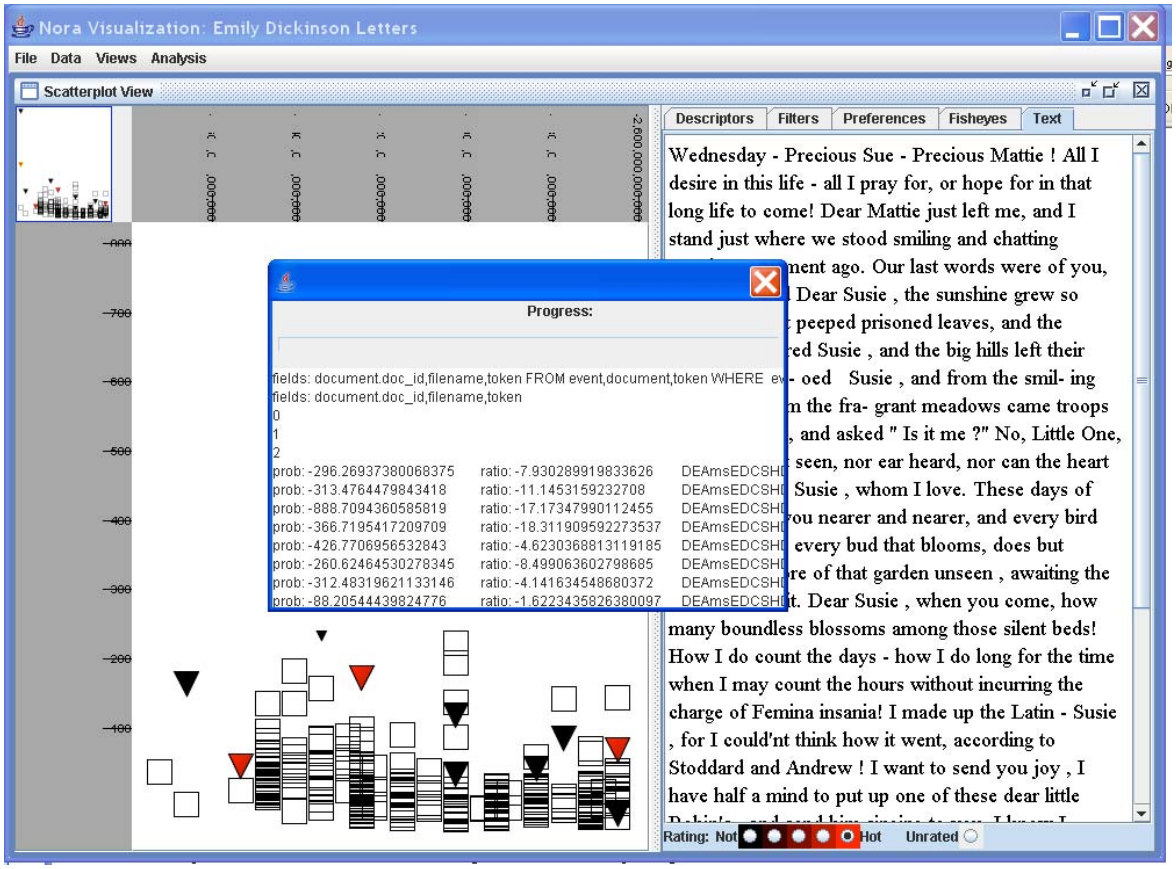

Fig. 6: D2K predicting the classification of the remaining letters.

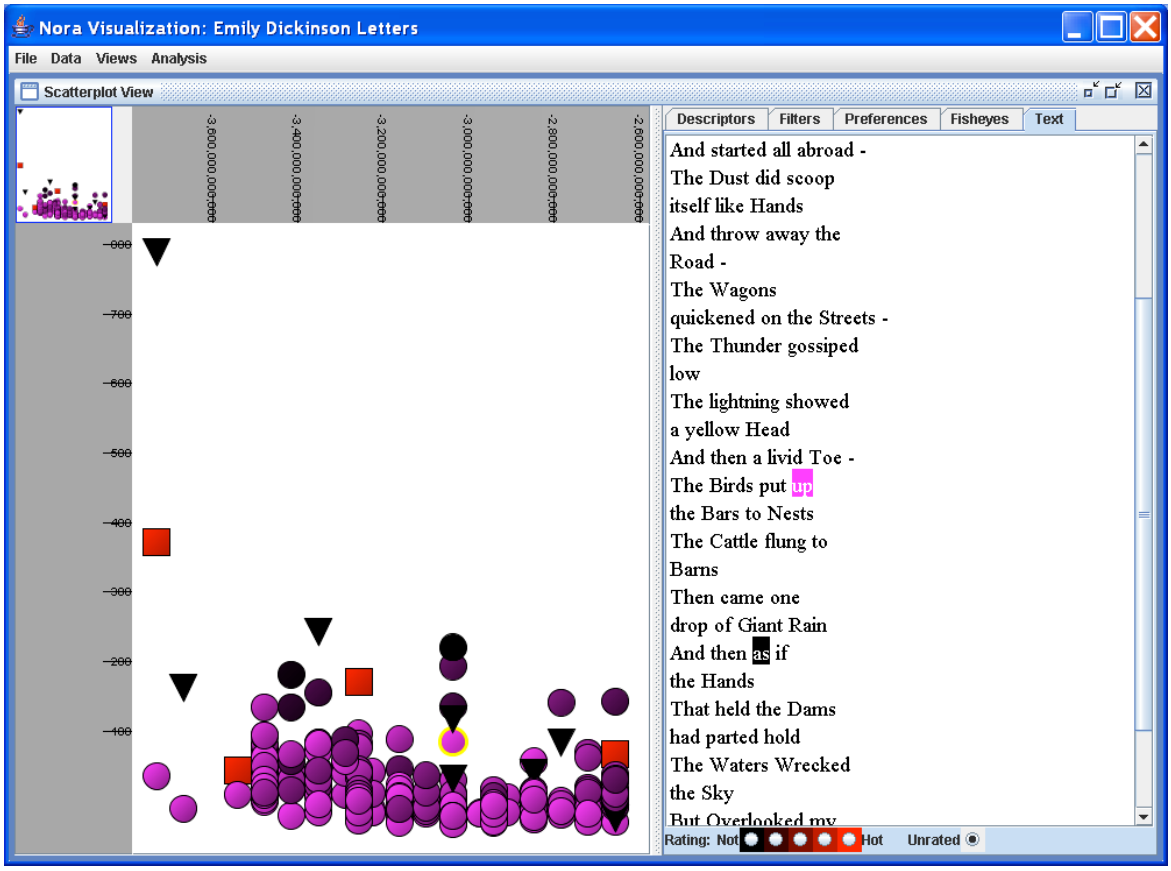

Fig. 7: All suggestions returned by D2K are shown as round dots, while the manually classified letters remain triangles.

The suggestions are purple when suggested to be hot, and black when not. One of the triangles became a square, indicating that there is disagreement between the manual classification and the D2K prediction for that letter. When users click on a dot to read the letter, the words that have been identified by D2K as good indicators for the classification are highlighted (again: purple for hot, and black for not hot). Users can change the display to make the more promising suggestions clearer by making the size of the items proportional to their prediction confidence, revealing the letters most likely to be of interest.

Reflecting on these experiments in an email to the project listserv, Martha Nell Smith (a scholar of Emily Dickinson) wrote:

> In the 1990s Harold Love stated something very important about "undiscovered public knowledge": 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. Five years ago I wrote about that fact in "Suppressing the Book of Susan in Emily Dickinson," an article surprisingly few Dickinson scholars seem to know, precisely because, though it's well situated in the volume Epistolary Histories, it is separated from much Dickinson criticism. Love does not remark anything we don't already know in some way, shape, form, and that is, I suppose, precisely the point. At one point or another members of this NORA team have been frustrated over the "oh wow" moment that just seemed to be missing. When my first one came, I was left saying, "uh, duh"—the "oh wow" moment is right in front of me/us. . . . . 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 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.
>
> The same goes for "write" – oh to leave a piece of oneself with, for, the beloved. To "write" is to present oneself, or a piece of oneself, physically – noting that the data mining was picking up both "write" when recorded by Dickinson and "write" in the header led the three of us to a "can we teach a computer to recognize tone" discussion. I wonder, remembering Dickinson's "A pen has so many inflections and a voice but one" what the human machine can do, what the human machine does, recognizing, identifying tone, what we think we're doing when we're so damned sure of ourselves.
>
> 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. On this list I've already talked about the limitations of "key words," a fact of which all humanists who get frustrated with search and retrieval are all too well aware, so I won't go on at great length about that. "Key words" are indispensable, but they don't work like magic, and we need to be rigorously self-conscious about all such taxonomies. I knew that, but it still surprised me when I saw texts that had several key erotic words and the texts were definitely not "hot."
>
> So Harold Love's observation very much holds—all of this was available to me but lay scattered as unrelated pieces. The data mining exercise was key to pulling it all together. Oh, and perhaps it goes without saying that the exercises also made me pull some things apart in order to make these connections.

We think this is a significant finding—that data-mining can provoke reflection and discovery in terms readily recognizable as traditional literary scholarship and criticism. We think it is important, moreover, that Martha Nell Smith's focus, above, is on the literary evidence, and not on the tool that uncovered it: that is a hopeful sign, in a project that means to facilitate the work of non-computational humanists.

Finally, at this writing Catherine Plaisant is taking the lead in submitting a paper on NORA for the JCDL conference in June in Chapel Hill, and Matt Kirschenbaum has organized the submission of a panel proposal on NORA for the Digital Humanities 2006 conference in Paris (formerly known as the ALLC/ACH annual joint conference), with participation from Greg Lord, Tanya Clement, and Martha Nell Smith at Maryland, as well as others at other NORA sites.

University of Illinois:

The University of Illinois, Urbana-Champaign (UIUC) is the lead site in the NORA project, and it is where the core data-mining software and expertise has been developed. It is also the place from which subawards and budget matters are managed, in the Graduate School of Library and Information Science (GSLIS), and overall project direction is provided by the grant's principal investigator, John Unsworth, through periodic conference calls (roughly one a month), email communication through the project listserv (webviz@prairienet.org), regular onsite meetings with UIUC participants (about once a week) and infrequent all-hands meetings, face to face (for example, in August of 2005, in Washington, DC, with meeting space provided by the Association of Research Libraries). Unsworth's role in the project has been to coordinate activities across the five NORA sites, set the development priorities for the various pieces of

NORA as they are built in these five different places, and make the practical decisions that balance the need for speedy progress toward compelling proof of concept against the need to build software that actually works and that can be used in more or less spontaneous experimentation as the project progresses.

In the first year of the NORA project, GSLIS procured and set up the server that runs the project's web server, stores the collected XML and SGML texts, and runs an instance of D2K server and of Tamarind (and the relational database it uses, Postgres). Maryland and UIUC have shared content management on the NORA site, with initial Web design having been done by Greg Lord at Maryland, and John Unsworth sharing responsibility with Greg on updating web content. GSLIS technical staff also established and maintain the project listserv, and they provide systems administration and backup on the NORA web server (www.noraproject.org), which resides in the GSLIS server room.

UIUC has in place subaward agreements with Maryland, Virginia, Georgia, and Alberta, and also keeps track of a separate budget within UIUC that funds the participation of the National Center for Supercomputing Applications (NCSA). Group travel to meetings is generally arranged and reimbursed through GSLIS, and equipment purchases are often run through GSLIS as well.

UIUC participants, in addition to Unsworth, include two doctoral students at GSLIS, Bei Yu and Xin Xiang, and Loretta Auvil, from NCSA's Automated Learning Group. Other staff at NCSA have also contributed from time to time, and other faculty and graduate students at UIUC have occasionally engaged in meetings and discussions of the NORA project.

Xin Xiang came to UIUC in 2004, with an interest in Java and XML, and with background in statistics and computer science. His initial contribution to the NORA project was to work with libraries and scholarly projects to collect the XML and SGML texts, and to put them online in a password-protected site. Next, he worked with NCSA to get NORA participants access to the NCSA CVS where source code for D2K is maintained and updated. More recently, he has focused on producing Java modules for use with Tamarind or D2K, and to some extent on experimentation with XML indexing software for subsetting of collections. For example, Xin was able to create a Tamarind library that connects to the PostgreSQL database and converts raw data to a sparse matrix. This standalone JAR file combines three D2K modules written by Bei Yu, Xin, and staff at NCSA. Xin has also been working with John Cuadrado and others at the National Institute for Liberal Education, to bring their work on automated document clustering algorithm into D2K as a module that could be deployed in a D2K data-mining itinerary. The implementation of this algorithm in D2K is still being fine-tuned.

Bei Yu is currently writing a dissertation entitled "An Evaluation of Text Classification Methods for Literary Study," the proposal for which can be found on the NORA web site under "Publications and Reports" or at http://www.noraproject.org/beiyu-proposal-Jan04-2006.pdf . In this project, Bei will be comparing the impact of various algorithms and pre-processing choices on the performance of supervised learning processes—specifically, the Dickinson erotic poetry experiment, the sentimental novel experiment, and a third experiment to see whether text classification software can reliably distinguish between fiction and non-fiction across the entire NORA testbed. Bei has facilitated text-mining experiments at both Maryland and Virginia by helping to manually preprocess data when the NORA architecture was not ready to perform the task itself, and by using text-mining techniques to explore conceptual and vocabulary overlap in the literature of literary scholarship and of knowledge discovery, in order to suggest ways in which text-mining software could support, extend, or facilitate work that literary scholars were already doing in non-computational ways. Her research in this area strongly suggests that

classification, clustering, association, and genre studies are all actively conducted by literary scholars and reported in the scholarly literature, and in many cases the objects of study are such that text-mining software could indeed give scholars doing this kind of work greater reach across collections and more solid evidence from individual texts.

Loretta Auvil, in NCSA's Automated Learning Group, has been the NORA project's main contact within the group that develops and maintains the D2K data-mining software. She has assisted the project with set-up and operation of a D2K server (the one currently used in the live demo referenced under "Architecture," above), with consultation on D2K software and performance issues, and with consultation on database and performance issues with Tamarind. She has also acted as a liaison to bring others to the table when specialized expertise was required from others at NCSA, and she has regularly attended all-hands meetings, conference calls, and face-to-face meetings on campus at UIUC, as well as organizing one meeting that used the NCSA Access Grid to facilitate a distributed group discussion of NORA's software architecture.

Presentations on the work underway in the NORA project have been made at ACH/ALLC 2005 by Bei Yu and John Unsworth, and in the November 2005 Lyman Award lecture at the National Humanities Center. Loretta Auvil and Bei Yu have also put in proposals for papers to be delivered at Digital Humanities 2006.

University of Virginia:

Project staff from the University of Virginia's Institute for Advanced Technology in the Humanities worked with Tom Horton (Computer Science, UVa) to select an appropriate software system to support the project's wiki, and subsequently configured and maintained the NORA project wiki, which you can see at:

http://www.iath.virginia.edu/nora/wiki/pmwiki.php?n=Main.HomePage
(password on request)

The wiki's navigational sidebar has links to topics of ongoing interest, for example:

- Important Dates
- Meeting Notes
- An Archive of Webviz, the project email list
- phpPgAdmin for Tamarind (an administrative interface to the RDMS)
- A Glossary
- VizGallery: interesting examples of information visualization.

In addition, the home page of the wiki links to ongoing work in sub-projects, for example:

Requirements Info:

- Workflow Description
- Terms and Definitions for User Needs & Requirements
- Open Questions
- Document Elements
- What Humanists Do

Software Design and Tools:

- Tamarind
- T2K using D2K
- WebDesign

- Architecture Questions
- End-to-End Demo
- Visualization Tool Design
- Whiteboard View of How NORA will work
- Recommended Tools we could use

Emily Dickinson Demo

- What We Have Done (in brief)
- Emily Dickinson Erotic Language
- Emily Dickinson: Hot and Not
- Try This to test the interface, with live connection to D2K

Sentimentalism Demo

- InASentimentalMood
- Experiment Plan
- Uncle Tom's Cabin Scoring
- Sentimental Rubric
- Tom and Ben's Workspace

Two UVa graduate students were brought onto the project in May 2005:

- Kristen Taylor, a PhD student in English, with interests and skills in 20th century literature, web design, and media studies.
- Ben Taitelbaum, a PhD student in Computer Science, with skills and interests in open-source development, software engineering, and computer graphics.

During the summer and early fall of 2005, project efforts at the University of Virginia centered on the following areas:

A study of sentimentalism:

An early goal was to define a problem area in literary studies that might lend itself to study using data-mining methods and that had other important characteristics:

- The problem could be studied using texts in the NORA testbed collection.
- The problem would be of real interest to literary scholars.
- There would be scholars at Virginia with a strong interest in this problem area.
- The problem would involve exploration at a different level of granularity within the NORA testbed, as compared with Maryland's Dickinson demo.

After brain-storming with members of the NORA project at other universities, we chose to student sentimentalism in 19th-century American novels. The term "sentimental" is first applied to 18th-century texts such as Henry Mackenzie's *The Man of Feeling* and Samuel Richardson's *Pamela*. Sentimental novels emphasize, as does Mackenzie's title, men and women of feeling, and feeling is valued over reason.

This prototype is still in process, with some parts completed and others not. When complete, it will use a set of texts in Virginia's Early American Fiction (EAF) collection (developed with support from the Mellon Foundation) that regularly appear on the syllabi of courses on American Sentimentalism.

Kristen Taylor came to the NORA project with an interest in this subject area, and together with Tom Horton, she has developed a plan for a series of experiments to examine this question. Kristen and Tom met with two scholars in Virginia's department of English (Marion Rust, who specializes in British literature, and Stephen Railton, who specializes in American literature) to discuss those plans and also to gather information about how they use texts in the EAF digital

library. They used this opportunity to have a more general discussion of how software might support recognized needs in literary research and teaching. Follow-up meetings are planned, once the prototype is operational.

Our plan for a series of staged experiments is fully described on the NORA wiki, at

http://www.iath.virginia.edu/nora/wiki/pmwiki.php?n=Main.SentExperimentPlan1

Those plans are more briefly summarized below.

Stage 1: Classification Using a Core Set of Novels

The initial goal is to evaluate the use of text-mining on a small set of "core" sentimental texts. These texts are considered strongly sentimental and are frequently taught as exemplars of this genre. The three novels we chose are:

- *Charlotte: A Tale of Truth* (two volumes, 35 chapters), by Susanna Rowson.
- *Uncle Tom's Cabin* (two volumes, 45 chapters), by Harriet Beecher Stowe
- *Incidents in the Life of a Slave Girl* (41 sections), by Harriet Jacob

Significant efforts since the summer have focused on assigning a ranking for "level of sentimentalism" to each chapter in these works. Assigning category labels for a training set is an important first step in any supervised learning process, but in literary study there is no objective value one can select to measure something like the level of sentimentalism: an individual scholar's judgment is required. But will more than one scholar label a section of text (e.g. a chapter) the same way? Can a group of scholars explain their decision criteria and reach consensus on a common rubric?

We devised a small experiment to test this inter-rater reliability, where eight scholars ranked the sentimentalism of three chapters from *Uncle Tom's Cabin* on a scale of 1-10. Results were remarkably consistent, especially when we grouped values into three categories: high, medium and low. From reading each person's explanation of his or her ratings, we believed it would be possible to develop a common set of criteria that could be used in scoring more chapters. (More details of this can be found on the wiki under the title "Uncle Tom's Cabin Scoring.") However, as the experiment has proceeded, we have found that scholars are not always so consistent. We hired four more graduate students to assign levels of sentimentalism to all 121 chapters in the three novels to be used in this stage. Each chapter was scored for sentimentalism on a scale of 1-10, by two students working independently. When these values were interpreted as high/medium/low there were 53 chapters where there was disagreement about rating. Very few of these were "high/low" disagreements: most were high/middle or low/middle disagreements, and these are currently being reassessed. In the meantime, the average numeric score has been converted to a high/medium/low score. We believe that this preliminary work has led to a better understanding of how sentimentalism may (or may not) be recognized by literary scholars. Kristen Taylor is considering assessing this as an outcome that might interest literary scholars in its own right.

In the meantime, as of late October 2005 the average-score labels are being used to begin data-mining studies with assistance from Bei Yu and others at UIUC. From the basic kinds of data mining that the NORA team has carried out on other collections, we believe we will quickly be able to:

- Evaluate whether D2K can classify chapters by level of sentimentalism in a small, well-understood collection.

- Determine the size of training set needed to succeed in classifying test-set chapters that are similar to the training set (i.e. from the same novels).
- Discover what word-occurrence information D2K finds significant when classifying by sentimentalism.
- Ask whether these results provide any insights that interest literary scholars who study these works.

We also plan to treat the chapters where our scholars assigned different high/medium/low scores as if they were of unknown classification, and see how D2K classifies them based on the set of chapters for which the scholars agreed.

Stages 2, 3 and 4 of the Study:

Stages 2, 3 and 4 of our experiment build on Stage 1's results. In Stage 2, we will add to our experiment several more works by the same authors. Our goal here is to learn how the data-mining classifiers perform when presented with data very similar to the previous training set. In Stage 3, we will add as many as seven new novels to the mix. Our goal in this stage is to evaluate previously successful strategies with a larger set of novels. The texts added will be those that scholars recognize as exhibiting sentimentalism, though some may not be as consistently sentimental (chapter-by-chapter) as the "core" set used earlier. Finally, in Stage 4, we will use text-mining on a set of works that are considered by scholars not to be wholly sentimental or not sentimental at all. At least five additional novels from the EAF or similar sources will be used in addition to the previous works. One of our goals in this stage is to identify parts of texts that contain aspects of sentimentalism, or common word-use that is sentimental in one context but not in another.

Software Architecture:

Tom Horton and Ben Taitelbaum have begun work that could lead to a more robust software architecture for applications being developed for the project. Our near-term goals are to work more closely with other NORA sites to see if our design model can support their needs and, if so, to help them modify existing applications to make use of our design models.

Areas of work in this area include the following:

- Defining a complete and robust interface between end-user applications and remote "NORA services" i.e. the applications running on the NORA server that carry out data mining and that return information in data-stores created by Tamarind (or other data stores).
- Defining a model for identifying and using documents organized into collections (either by the way they are ingested into Tamarind or at run-time by the end-user).
- Defining a model and mechanism for assigning properties or attributes to documents, parts of documents, and other data entities of interest to NORA tool developers and users.
- Developing a method of allowing the retrieval of parts of documents based on XPath information stored in Tamarind.

As part of his regular duties as a computer science professor, Horton is required to help a small number of fourth-year students identify and carry out research projects for their required senior thesis. This year, he actively recruited students to work on small projects involving software tools for literary studies. Experience shows that not all such students complete work in a timely fashion or produce useful results, so he has decided not to engage these students in projects that were central to the NORA project plans. But there are three students working on projects closely related to NORA project work:

- Christen Haden is working on a visualization tool that will use the Tilebar method to display numeric data (such as word token counts etc.) that might normally be displayed in a table. The relation to the NORA project is that we will assume that the display of data by level of "segmentation" within a document, e.g. by whole document, by chapter, by page, etc. Her application is being developed in a way that information from Tamarind could be used as the source of the data. Christen's work will be completed by December 2005.
- Ryan Ahearn is working on advanced search methods that better meet the needs identified by the two English scholars interviewed this past summer. He will develop approaches that make use of neighborhood searches, searches on documents found by previous results, etc. He is also interested in visualization methods to show strength of occurrence by region in a document. He will apply this work to the novels used in the first two stages of the sentimentalism experiment. His work will be completed by April 2006.
- Min Han will explore using results generated by semantic tagging tools developed by Lancaster University's University Centre for Computer Corpus Research on Language (UCREL). Their USAS system (http://www.comp.lancs.ac.uk/ucrel/usas/) assigns a semantic tag-value to every word-token in a document, making use of a semantic tag-set defined by corpus linguists. Working in cooperation with contacts at Lancaster, we will explore using their tools to tag novels used in our sentimentalism study. If this is at all successful, then semantic tags could be used in data-mining studies. In the meantime, Min will make use of a previously semantically tagged version of Shakespeare's plays to explore how searches based on semantic values and word-types can be utilized.

These projects may bear fruit that will be useful to the NORA project. If so, then the work will be continued with NORA project funds.

6. Reference URLs

The NORA Wiki (password on request):
http://www.iath.virginia.edu/nora/wiki/pmwiki.php?n=Main.HomePage
This is the private working site for project participants.

The NORA Web:
http://www.noraproject.org
This is the public face of NORA.

A Live Demo of NORA:
http://www.noraproject.org/demos.php

The Webviz Archive (password on request):
https://mail.prairienet.org/mailman/private/webviz/

Publications and Reports by NORA project participants:
http://www.noraproject.org/publications.php