Supporting Literary Scholars with Data Mining and Visual Interfaces

James Rose, Catherine Plaisant
HCIL

Matthew G. Kirschenbaum, Martha Nell Smith, Tanya Clement, Greg Lord
Maryland Institute of Technology for the Humanities and Dept. of English

Bei Yu*, Loretta Auvil^
University of Illinois (*GSLIS and ^NCSA)
Goal

- Support literary scholars

- Visual interface for data mining
  - Accessible
  - Provocational
Context

• **Text Mining - Machine Learning**
  • Industry, defense, science, law enforcement
  • Humanities: just beginning!

• **Nora Project**
  • [www.noraproject.org](http://www.noraproject.org)
  • Mellon Foundation
  • U. of Illinois, Georgia, Virginia, Alberta and Maryland
  • 18th and 19th Century British and American Literature
  • Work presented here:
    • Maryland (HCI and literary research) and Illinois (Data Mining)
The Valley Project details life in two American communities, one Northern and one Southern, from the time of John Brown's Raid through the era of Reconstruction. In this digital archive you may explore thousands of original letters and diaries, newspapers and speeches, census and church records, left by men and women in Augusta County, Virginia, and Franklin County, Pennsylvania. Giving voice to hundreds of individual people, the Valley Project tells forgotten stories of life during the era of the Civil War.

Enter the Valley Archive

Copyright 1993-2006, All Rights Reserved Edward L. Ayers
Women Writers Project

The Brown University Women Writers Project has its intellectual roots in two communities whose synergy began to be evident at the end of the 1980s. The first of these was the growing field of early modern women's studies, whose project was to reclaim the cultural importance of early women's writing and bring it back into our modern field of vision. The other was the newly developing area of electronic text encoding, with its emphasis on improved access and long-term preservation of textual data. As a method of bringing inaccessible texts back into use, the electronic archive seemed like the ideal successor to the physical archive, since it promised to overcome the problems of inaccessibility and scarcity which had rendered women's writing invisible for so long. This partnership of archival scholarship and electronic technology has become a model for text encoding projects all over the world. In our newsletter we publish excerpts from the textbase, and articles on issues in text encoding and women's writing.

In the first five years of the project, we transcribed an initial collection of about 200 texts, and began making draft printouts available for teaching and research. These printouts are still available through our online ordering system, and we will be issuing new, updated versions soon. We also worked on a project with Oxford University Press to publish editions of selected texts in traditional print form.

In 1993, with the publication of the expanded TEI Guidelines, the WWP began a three-year period of research on how to use the new guidelines for early women's texts, and how to convert our existing encoding to the new model. During this interval, we encoded very few new texts, but we established a new set of encoding methods, set up improved systems of documentation and training, and began the long process of converting our legacy data.

With the new encoding system in place, we resumed encoding texts in earnest in 1996. From 1997 to 2000, with a grant from The Andrew W. Mellon Foundation, we developed Renaissance Women Online, a project studying the impact of electronic texts on teaching and research. With support from the Rhode Island Committee for the Humanities we also sponsored "In Her Own Words", a one-woman show based on the life and writing of Elizabeth I. And the National Endowment for the Humanities renewed our funding once more in 1999 to encode a group of new texts focusing on satire, gender politics, and the cultural context of 18th-century England.
Literary Text Analysis

- New area
- Many classification tasks
  - Authorship attribution
  - Stylistic analysis
  - Genre Analysis, etc.
- Scholars need assistance to operate
Text Visualization Examples

- Compus (structural XML components) [Fekete 00]
- SoftVis (line attributes) [Eick, 92]
- Text Arc (frequently occurring words, circular overview) [Paley, 02]
- Themeview (topic overview) [Wise 95]
Our Users' Needs

- Situate: show what is available
- Characterize*: review characteristics of docs in a collection and techniques available
- Read*: most essential activity
- Classify*
- Find indicators*
  - What makes a doc fall in a category or another?
- Understand results*
- Interpret: annotate, refer, illustrate etc.
- Compare
- Archive and disseminate

(* tasks with started to address)
Our users

- Small group of enthusiastic computer users
- Broad base of scholars uninterested in computational tools
  (unless they offer simple and comprehensive user interfaces)

- Argumentation and provocation very valuable
  (in contrast to other domains that focus on decision making)
1st selected case study

- Collection of Emily Dickinson’s letters
  - About 300 XML encoded documents
  - Correspondence with Susan Huntington Dickinson (her sister in law)
- Study:
  - Patterns of erotic language
  - Controversial topic
  - Work with domain experts
Definition

- A document rated as "erotic"
  - is flirtatious
  - has sexual connotations
  - is seductive, enticing
  - aims to pull the addressee in by imbuing the emotional and the intellectual with attention to the physical, even physical arousal.
- the text thus intends to be sexual or sexualized in some way(s)
- we use the short word “hot”
Question: Does this document show signs of erotic language?  
False (i.e. not hot)     True (i.e. hot)
A little bread, a crust - a crumb,
A little trust, a Demijohn,
Can keep the soul alive,
Not portly - mind!
But breathing - warm -
Conscious, as old Napoleon
The night before the Crown!
A modest lot, a fame petite,
A brief campaign of sting and sweet,
Is plenty! is enough!
A sailor's business is the Shore -
A soldier's - Balls!
Who asketh more
Must seek the neighboring life!
Dear Sue,

God bless you for the Bread!
Now - can you spare it?
Shall I send it back?
Will you have a Loaf of mine - which is spread?
Was silly eno' to cut six, and have three left.
Tell me just as it is, and I'll send home your's, or a Loaf of mine spread, you understand - Great times - Love for Fanny.
Wish Pope to Rome - that's all.

Emily.

Escr. in parlor -
<table>
<thead>
<tr>
<th>ID</th>
<th>Prob</th>
<th>title</th>
<th>you</th>
<th>heart</th>
<th>sister</th>
<th>face</th>
<th>me</th>
</tr>
</thead>
<tbody>
<tr>
<td>231</td>
<td>0.0</td>
<td>Sue - to be / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>232</td>
<td>0.0</td>
<td>To be Susan ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>233</td>
<td>-5.6</td>
<td>Gratitude - / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>234</td>
<td>5.4</td>
<td>Sue - this / / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>235</td>
<td>2.5</td>
<td>Susan - / Th / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>236</td>
<td>0.0</td>
<td>Sweet Sue,/ ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>237</td>
<td>-1.0</td>
<td>We meet / / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>238</td>
<td>-3.0</td>
<td>To lose what / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>239</td>
<td>14.9</td>
<td>Dear Sue, / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>240</td>
<td>4.8</td>
<td>But Susan is ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>241</td>
<td>8.3</td>
<td>Susan - I won ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>242</td>
<td>3.7</td>
<td>Dear Sue is / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>243</td>
<td>7.4</td>
<td>Dear Sue / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>244</td>
<td>7.9</td>
<td>Susan is / / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>245</td>
<td>4.3</td>
<td>Do n't do suc / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>246</td>
<td>6.0</td>
<td>Susan's Idol / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>247</td>
<td>3.6</td>
<td>Memoirs of / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>248</td>
<td>-2.0</td>
<td>Write! Comm / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>249</td>
<td>1.3</td>
<td>Dear Susie - / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.0</td>
<td>Only Woman / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>251</td>
<td>4.1</td>
<td>Dear Sue / / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>252</td>
<td>17.6</td>
<td>Sister, / We / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>253</td>
<td>-0.5</td>
<td>Were not Da / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>254</td>
<td>11.0</td>
<td>Dear Sue / / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>-4.2</td>
<td>A Death blo / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>3.5</td>
<td>Sister spoke / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>257</td>
<td>5.6</td>
<td>Wish had / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>258</td>
<td>0.4</td>
<td>It was like / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>259</td>
<td>1.2</td>
<td>The Sallows / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>260</td>
<td>0.0</td>
<td>Morning / m / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>261</td>
<td>19.5</td>
<td>Dear Sue / / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>262</td>
<td>4.4</td>
<td>Thank Susa / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>263</td>
<td>-4.3</td>
<td>Where we / / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>264</td>
<td>3.6</td>
<td>Part to who ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>265</td>
<td>-0.3</td>
<td>The igomnion / / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>266</td>
<td>1.1</td>
<td>Dear Sue - / / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>267</td>
<td>-2.7</td>
<td>Trifles - like / ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mother and Vinnie send much love - they will be delighted to see you. My dearest love to Mat.

Sabbath Day.

I am sick today, dear Susie, and have not been to church. There has been a pleasant quiet, in which to think of you and I have not been sick enough that I cannot write to you. I love you as dearly, Susie, as when love first began, on the step at the front door, and under the Evergreens, and it breaks my heart sometimes, because I do not hear from you. I wrote you many days ago - I won't say many weeks, because it will look sadder so, and then I cannot write - but Susie, it troubles me.

I miss you, mourn for you, and walk the streets alone - often at night, beside, I fall asleep in tears, for your dear face, yet not one word comes back to me from that silent West. If it is finished, tell.

I asked Austin if he had any messages - he replied he -
Evaluation

• Naïve Bayes algorithm
  • Validated (i.e. better than classification baseline: majority class)

• User observations
  • Two studies: erotics and spirituality
  • Reading and reviewing documents was easy
  • Prediction seemed reasonable
  • Predictors less useful, with exceptions
  • Some questions are better suited than others
Conclusion

- Data mining interfaces can be designed to be accessible to literary scholars

- There is a rationale for provocational text mining in literary interpretation

- JCDL 2006 paper - Video in your bag - Demo this afternoon

- For more information and demonstration: www.noraproject.org

- Thanks to all the members of the Nora team
  in particular John Unsworth who leads this project, Steve Ramsay who provides the Tamarind system, and David Clutter, Greg Pape, and Andrew Shirk from NCSA who helped setup the web services

- Partial support for this work was provided by:
  - the Andrew Mellon Foundation and
  - the University of Maryland Libraries