WordSeer.berkeley.edu

A Sensemaking Environment for Literary Text

Aditi Muralidharan aditi@cs.berkeley.edu
Marti A. Hearst hearst@ischool.berkeley.edu

NATIONAL ENDOWMENT FOR THE Humanities

Digital Humanities

More and more source text in the humanities gets digitized every day, making it accessible to large scale computational analysis. By contrast, traditional methods of humanistic analysis are based on detailed arguments built upon on close readings of individual texts. How will the field adapt? How do we use statistics and text mining to answer humanistic questions?

Text Analysis in the Humanities

To date, text analysis systems for humanities scholars have focused on aiding interpretation. First, they apply some form of natural language processing to extract aggregate statistics about word usage, topics, named entities, and parts of speech. Second, they display the extracted information with visualizations like word clouds, node-and-link diagrams, and lists of word contexts. Such systems make patterns of style, form, and theme visible, and interpretable by people.

However, literature study is a form of sensemaking: a cycle of reading, interpretation, exploration and understanding. As useful as they are, current digital humanities text analysis systems leave the exploration and understanding part of the cycle unsupported.

A Sensemaking Challenge

Past studies of sensemaking have focused on decision-making or intelligence analysis tasks. In these domains, the objects of analysis are explicit and clearly defined: people, events, facts, relationships, and numbers. In literature study, by contrast, the objects of study are style, themes, imagery, form and stereotypes. How can text mining and visualization help literature scholars analyze these nebulous kinds of information?

Related Projects

The MONK project at CMU, and the Voyeur and TaPor project at McMaster University share the same cause as WordSeer. They use basic visualization and language processing to help literature scholars explore language use. The WordHoard program uses detailed metadata to give users tables of word frequencies in different subsets of text. The Prism project at the University of Virginia's Scholar's lab is a tool for "crowdsourcing interpretation". It visualizes the results of many users' interpretations of a text.

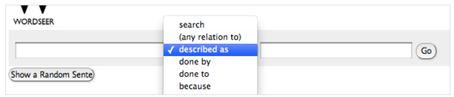
"How does the portrayal of men and women in Shakespeare change in different circumstances?"

We demonstrate WordSeer's capabilities by using it to explore this open-ended question. We find that when love is a major plot point, the language around women becomes more physical, for men, more sentimental.

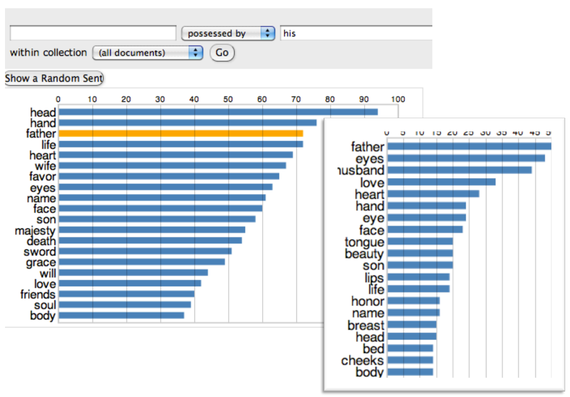
1 Grammatical Search

What are some things that are portrayed as 'his' and some things that are 'hers'?

Ordinary keyword search would return a list of sentence matches. The word **his** is always a possessive pronoun, so word sequences containing **his** would nearly always be relevant. But **her** can also be a 3rd-person pronoun, and will yield constructions like "I told her that X" and "I gave her the Y". With WordSeer, we make headway on this problem with grammatical search.



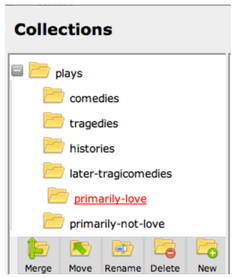
Above Using grammatical relationships extracted through natural language processing, users can query for relationships between words, such as "possessed by", "described as", "done to" and "done by". WordSeer users natural language parsers to extract this information.



Above Graphs show the frequencies of different words that are "possessed by" **him** and **her**. Body parts and male relatives dominate the picture for women.

2 Collections

Is the physical, patriarchal portrayal of women equally common in all types of plays?

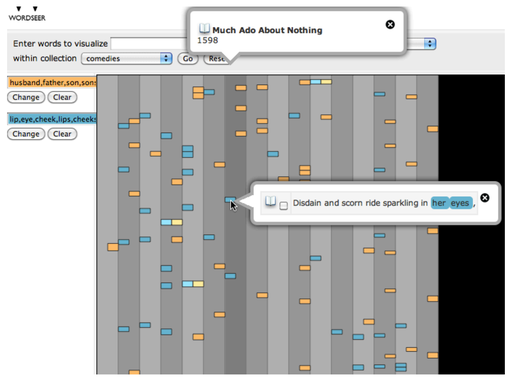


Left We use WordSeer's collections feature to create three initial sets: comedies, tragedies, and histories.

There is a document listing (not shown) from which plays can be added to sets.

3 Visualization

We use WordSeer's newspaper-strip visualization to examine the prevalence of "her body parts" and "her male relatives" in the comedies, tragedies, and histories.



Above each comedy is represented as a long column. Within each column, small, colored horizontal blocks (corresponding to 10 sentences each) highlight the presence of a match.

The results for the tragedies collection were similar to the results for comedies but in histories an interesting pattern emerged. It seemed that references to "her body parts" were somewhat less prevalent in the histories, but references to "her male relatives" remained common.

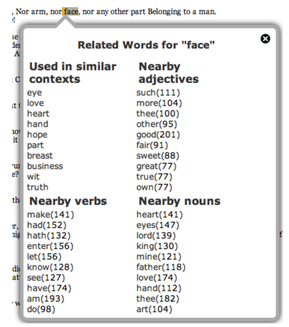
Hovering over a few body-part results in the visualization quickly led to a new hypothesis. In our rough sample, many of the mentions sounded romantic.

4 Exploration

Womens' body parts are mostly mentioned in romantic contexts

Above we use the reading and annotating interface to follow up on our hypothesis by clicking on the highlighted blocks in the newspaper-column visualization. We selected the speeches referring to body parts and tagged them by the topics they seemed to contain. It soon became apparent that many of the mentions were speeches by a lover.

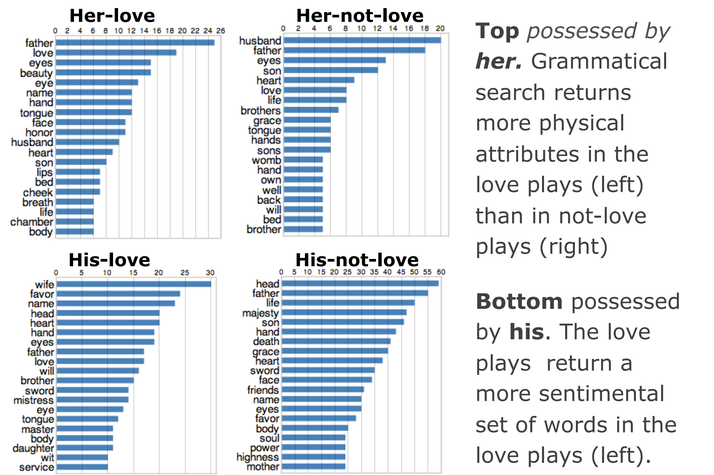
Left WordSeer uses distributional similarity to calculate similar words. Clicking on a word while reading brings up the option to see words used in similar contexts, and within 10 sentences. The related words for 'face' strengthened our hypothesis. They included other body parts, but also romantic words like 'love', `sweet', and `fair'.



5 Comparative Analysis

When love is a major theme, the language around women becomes more physical, and for men, more sentimental

We created a final pair of categories, "love plays", in which love was a major plot point, and "not love plays".



Top *possessed by* ***her.*** Grammatical search returns more physical attributes in the love plays (left) than in not-love plays (right)

Bottom possessed by **his**. The love plays return a more sentimental set of words in the love plays (left).