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

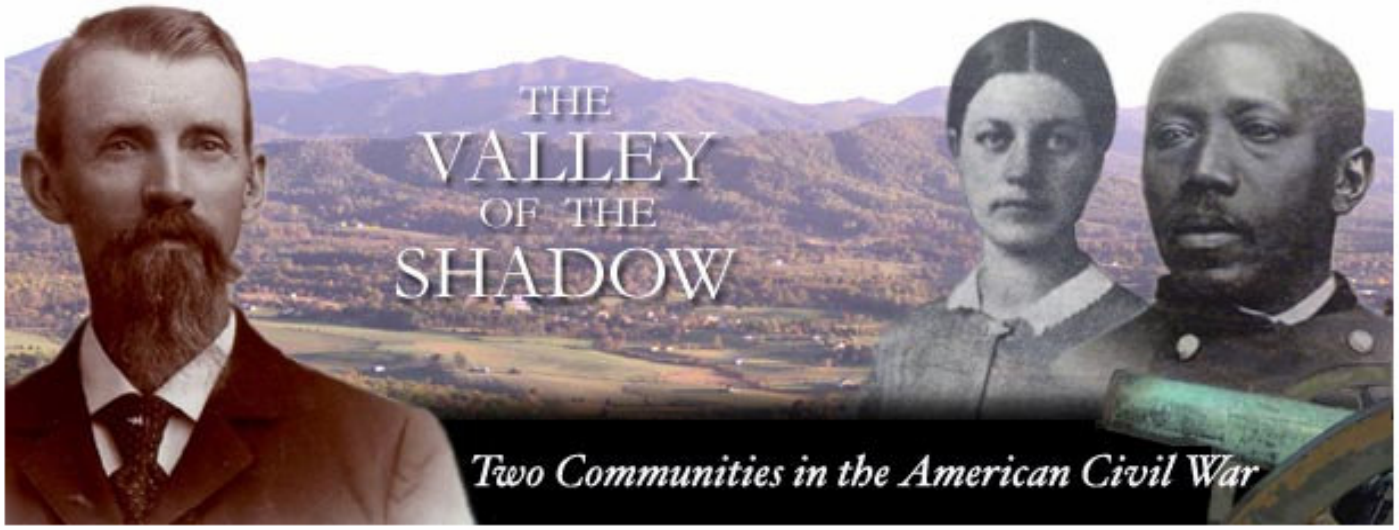
- Support literary scholars
- Visual interface for data mining
  - Accessible
  - Provocational

Context

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

- Nora Project
  - www.noraproject.org
  - 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)

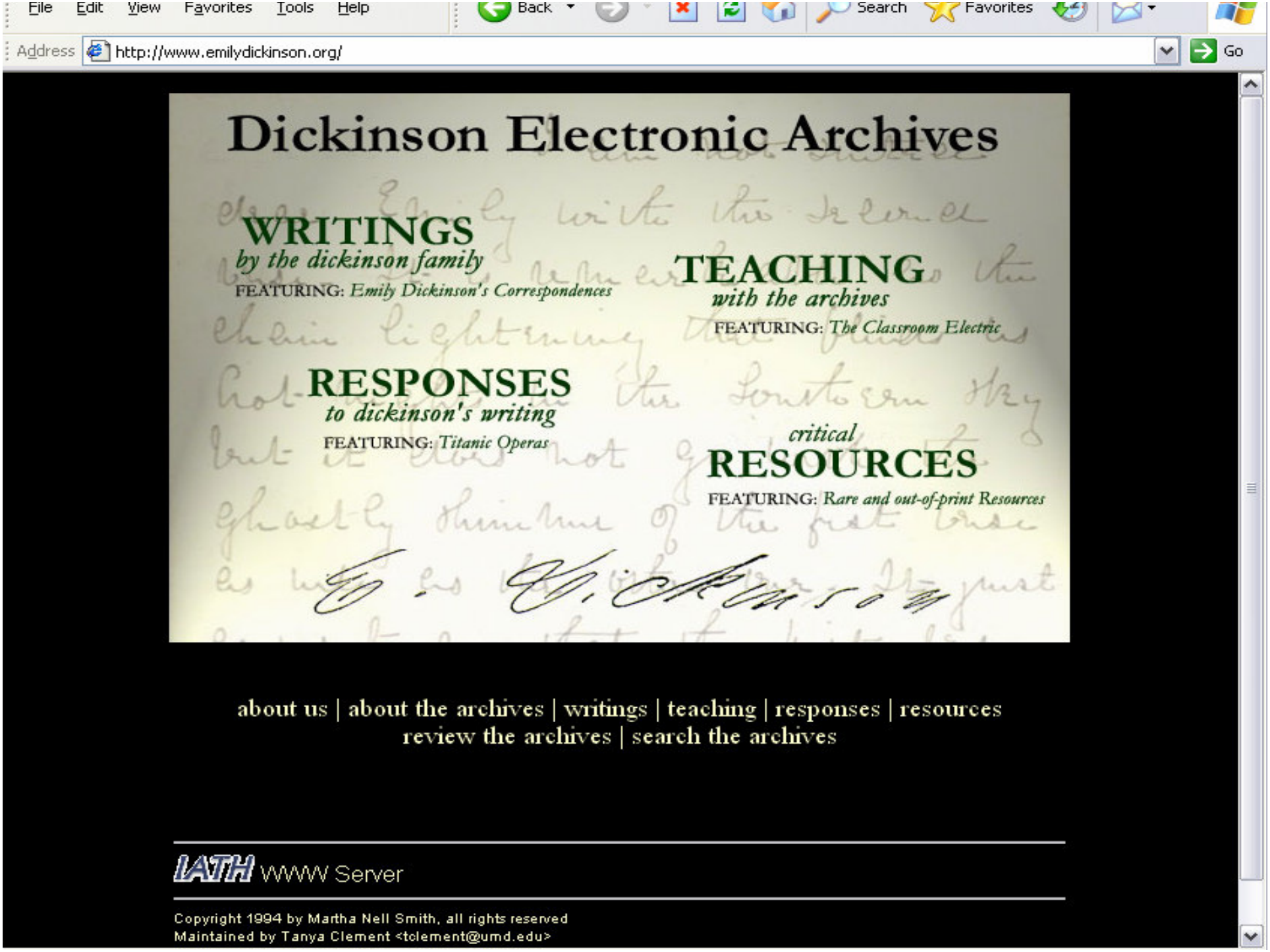
HCIL



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


File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.wwp.brown.edu/project/index.html
Go
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 longterm 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.

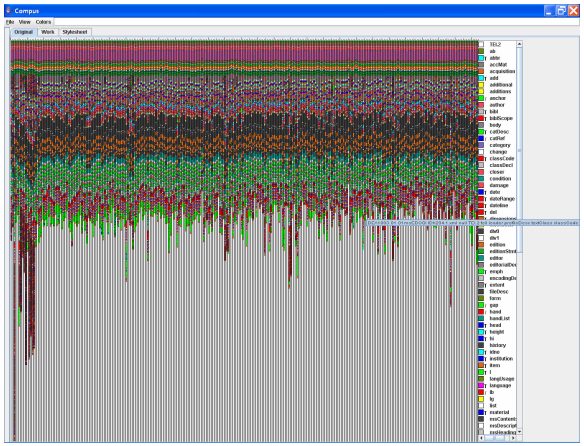

File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.emilydickinson.org/
Go
Dickinson Electronic Archives
WRITINGS
by the dickinson family
FEATURING: Emily Dickinson's Correspondences
TEACHING
with the archives
FEATURING: The Classroom Electric
RESPONSES
to dickinson's writing
FEATURING: Titanic Operas
critical
RESOURCES
FEATURING: Rare and out-of-print Resources
about us | about the archives | writings | teaching | responses | resources
review the archives | search the archives
IATH WWW Server
Copyright 1994 by Martha Nell Smith, all rights reserved
Maintained by Tanya Clement <tclement@umd.edu>

Literary Text Analysis

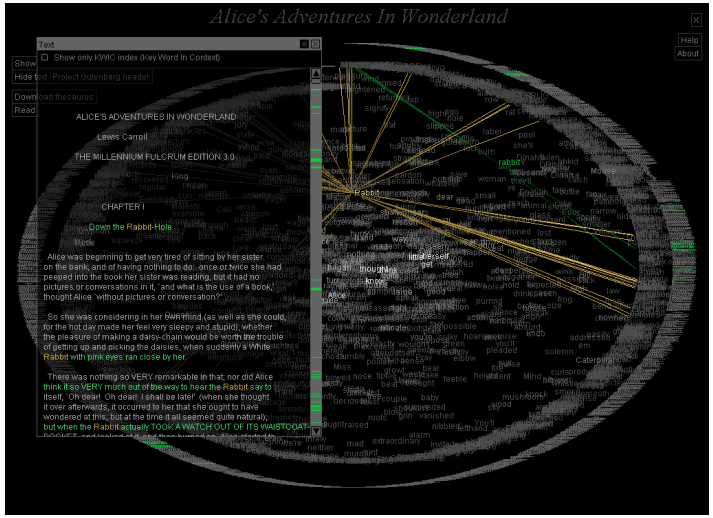
- New area
- Many classification tasks
  - Authorship attribution
  - Stylistic analysis
  - Genre Analysis, etc.
- Scholars need assistance to operate

HCIL

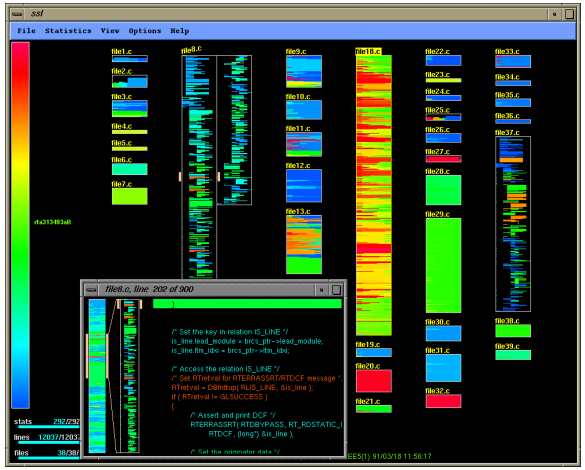
Text Visualization Examples



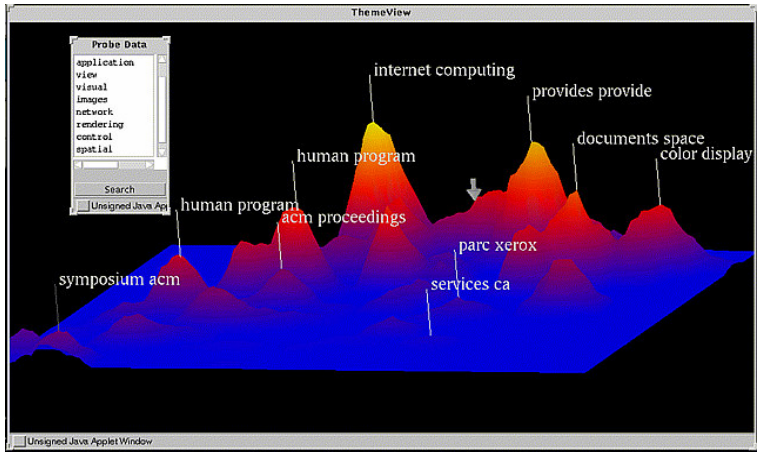
•Compus (structural XML components)
[Fekete 00]



•Text Arc (frequently occurring words, circular overview)
[Paley, 02]



•SoftVis (line attributes)
[Eick, 92]



•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 enthusiast computer users
- Broad base of scholars uninterested in computational tools

  (unless they offer simple and comprehensive user interfaces)

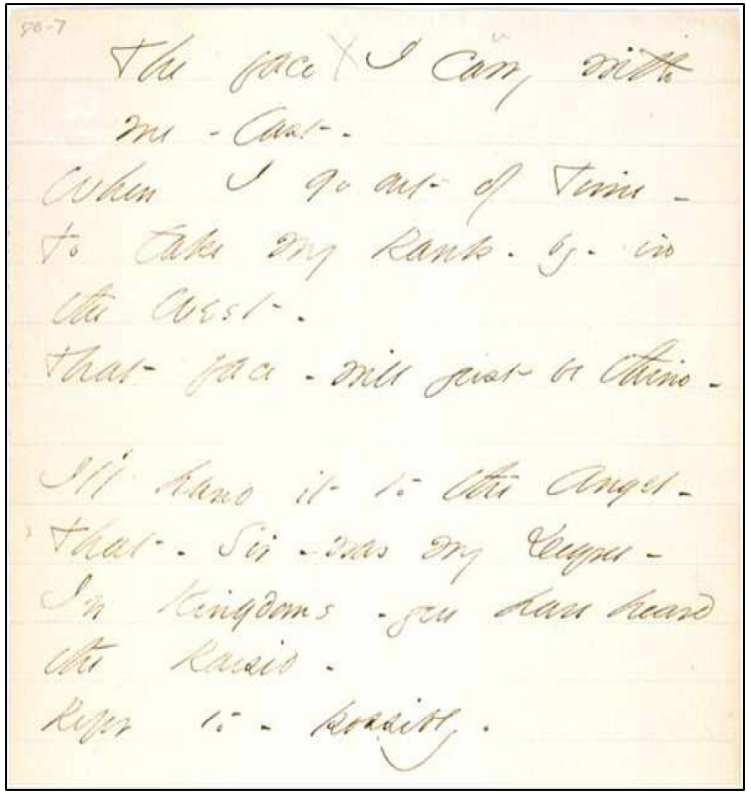
- Argumentation and provocation very valuable

  (in contrast to other domains that focus on decision making)

HCIL

1st selected case study

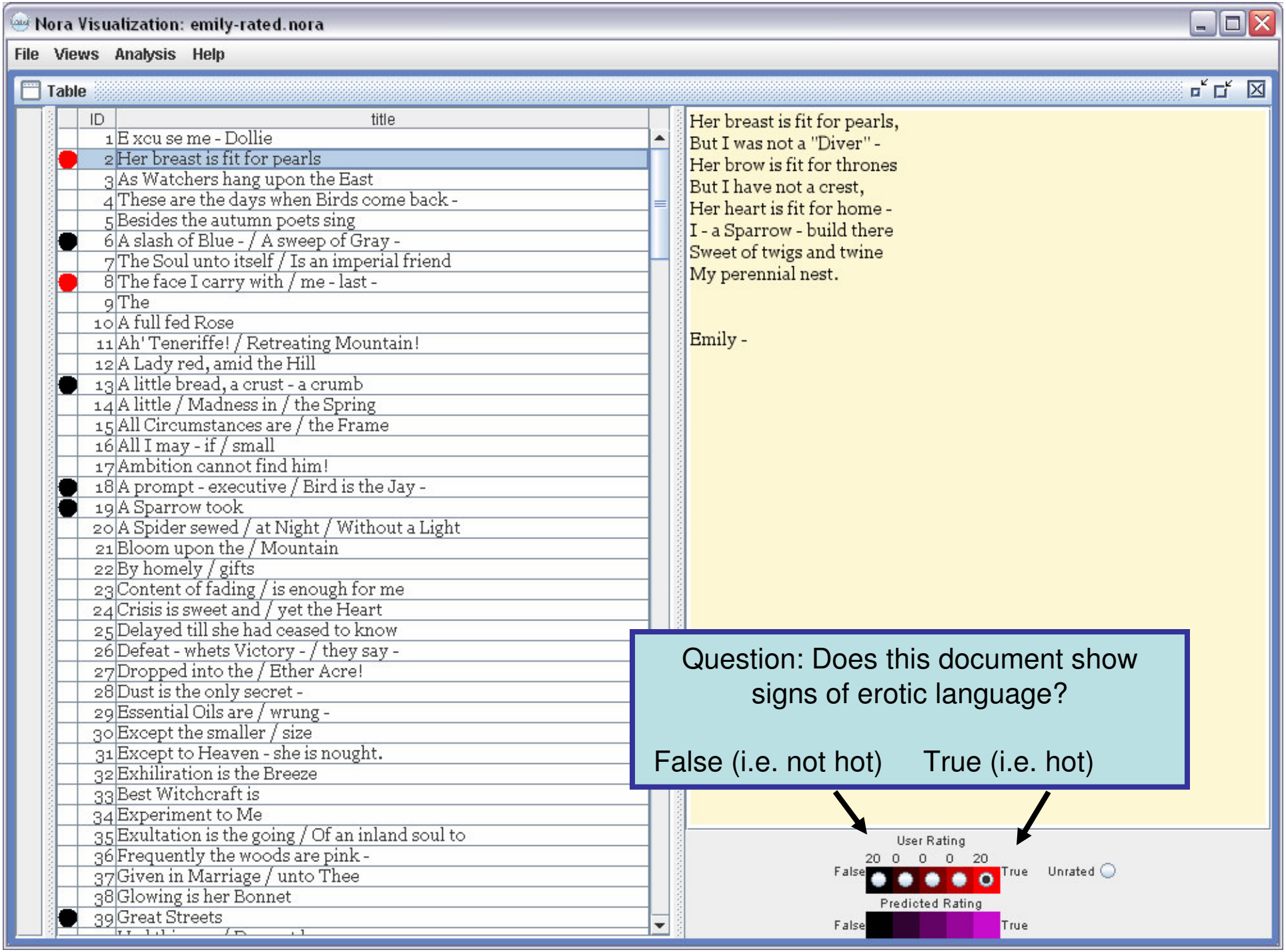
- Collection of Emily Dickinson's letters
  - About 300 XML encoded documents
  - Correspondence with Susan Huntington Dickinson (her sister in law)
- Study:
  - Patterns of erotic language
  - Controversial topic
  - Work with domain experts

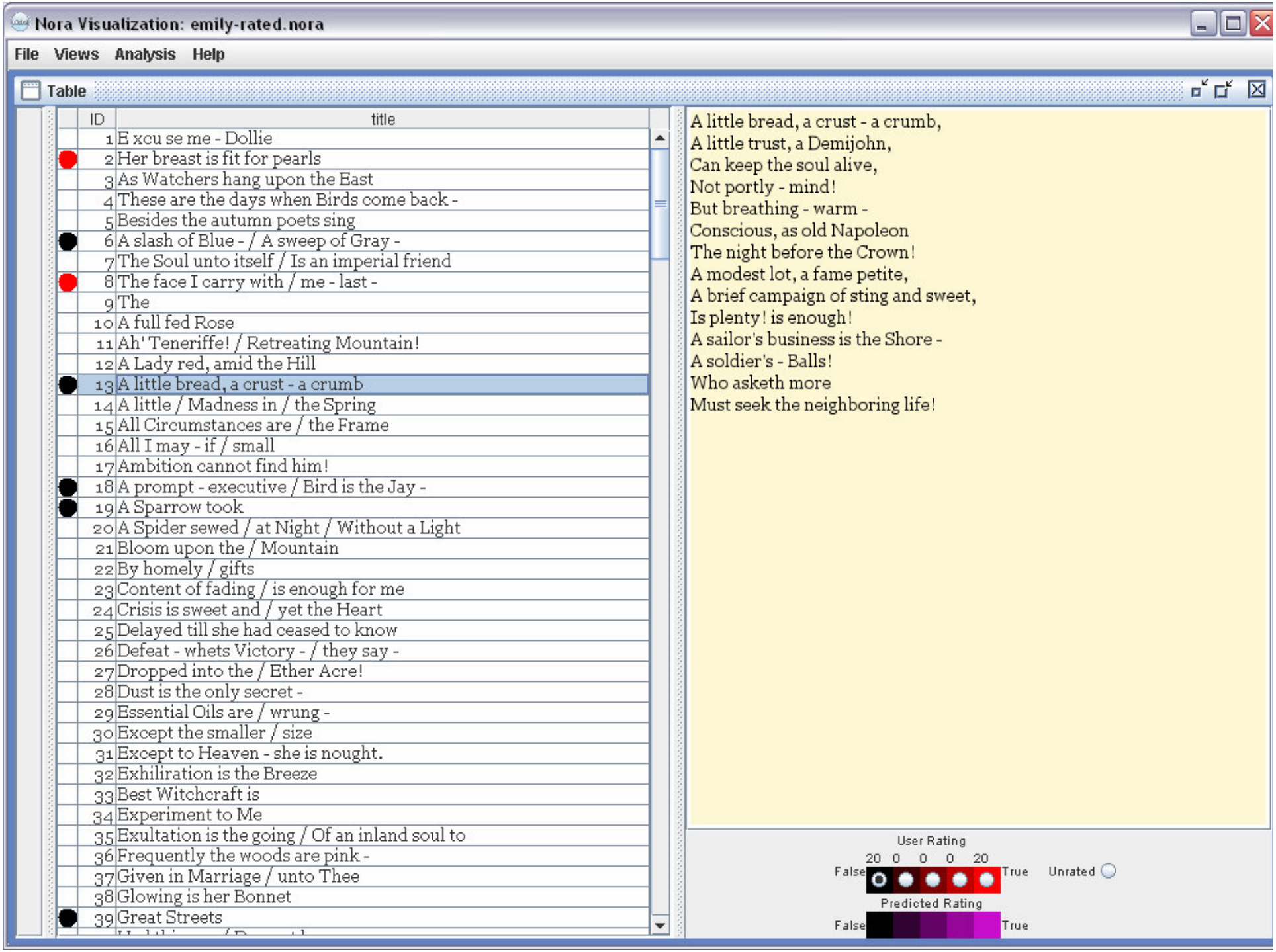


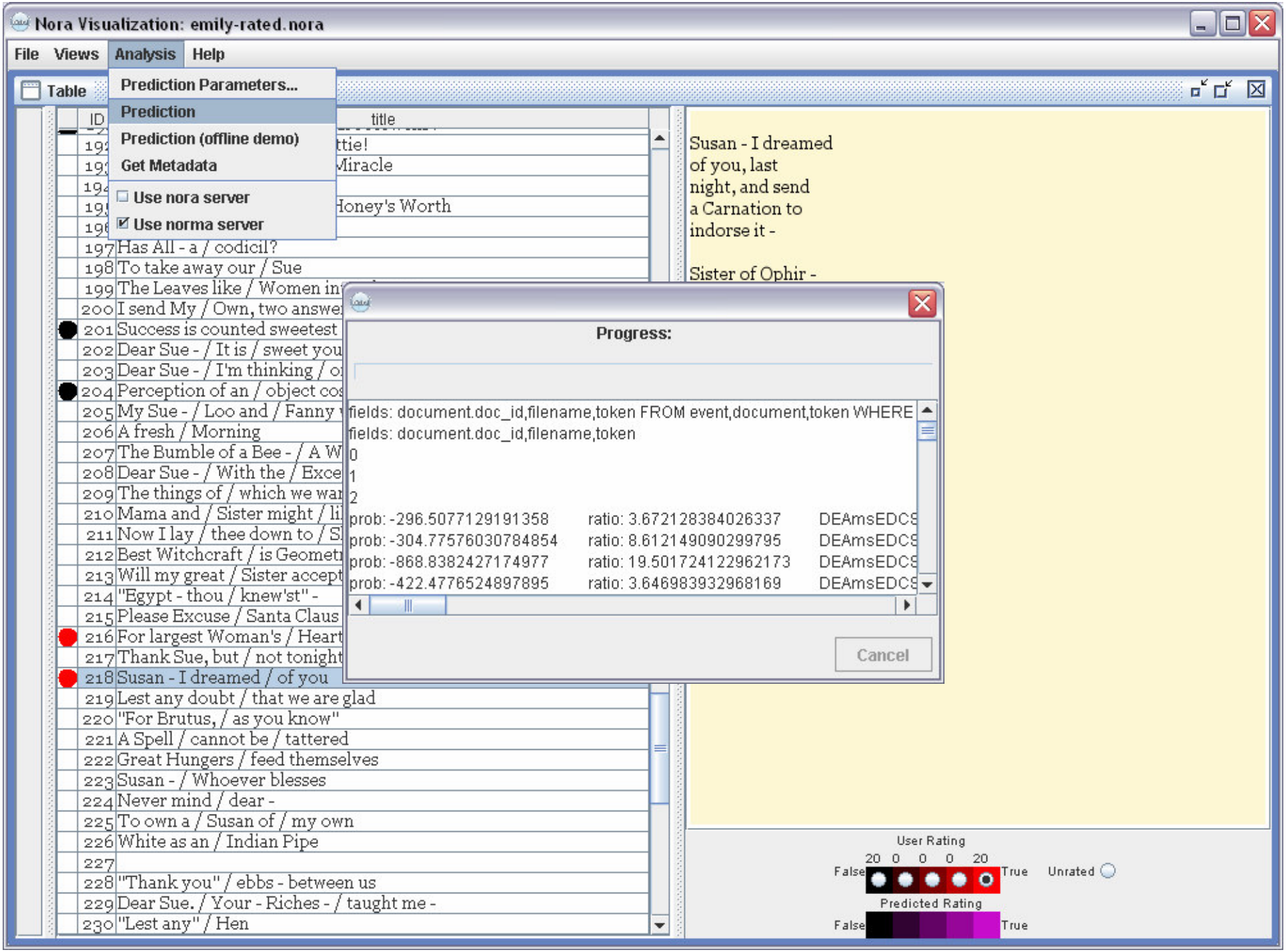
Definition

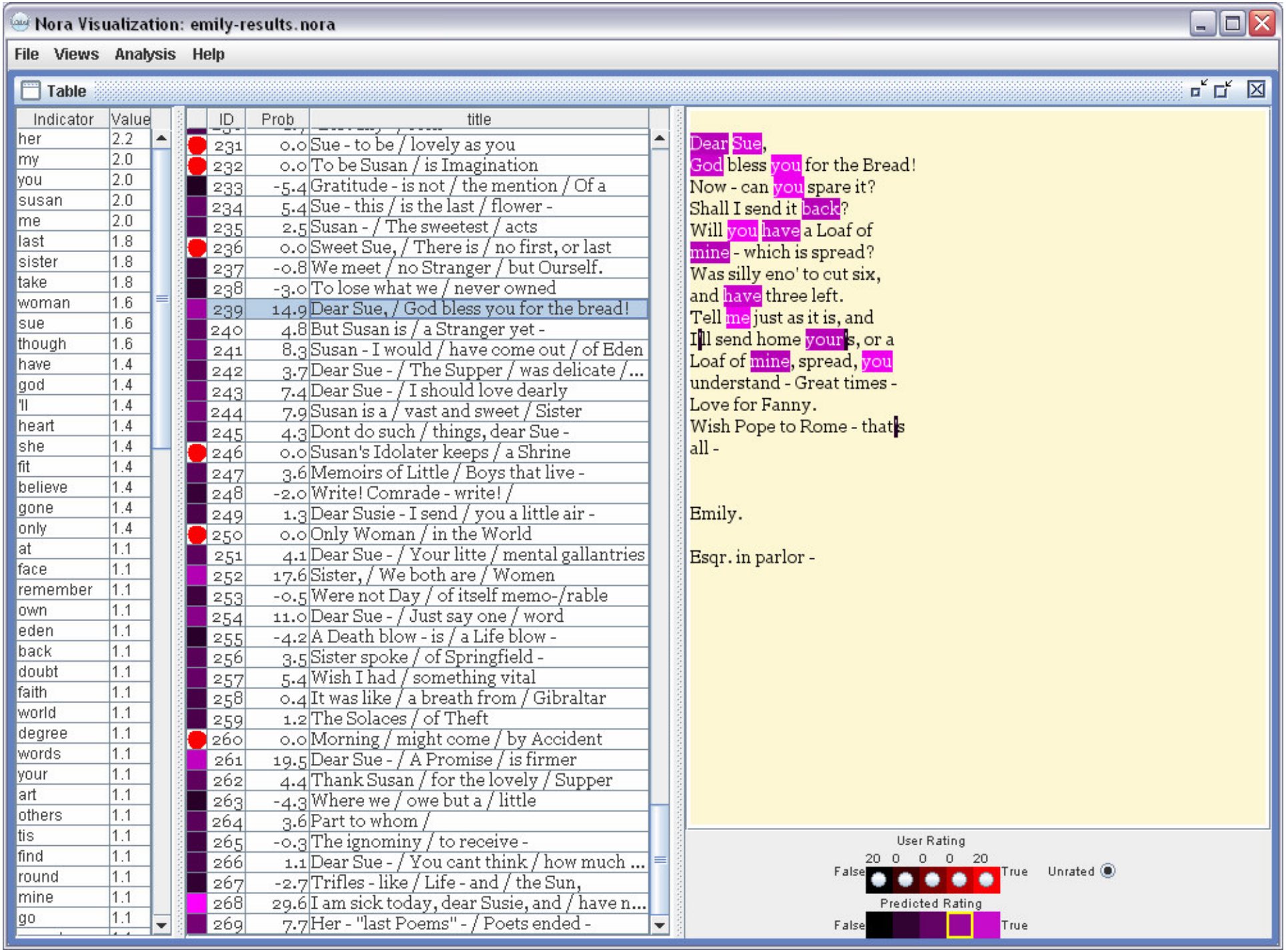
- A document rated as "erotic"
  - is flirtatious
  - has sexual connotations
  - is seductive, enticing
  - aims to pull the addressee in by imbuing the emotional and the intellectual with attention to the physical, even physical arousal.
- the text thus intends to be sexual or sexualized in some way(s)

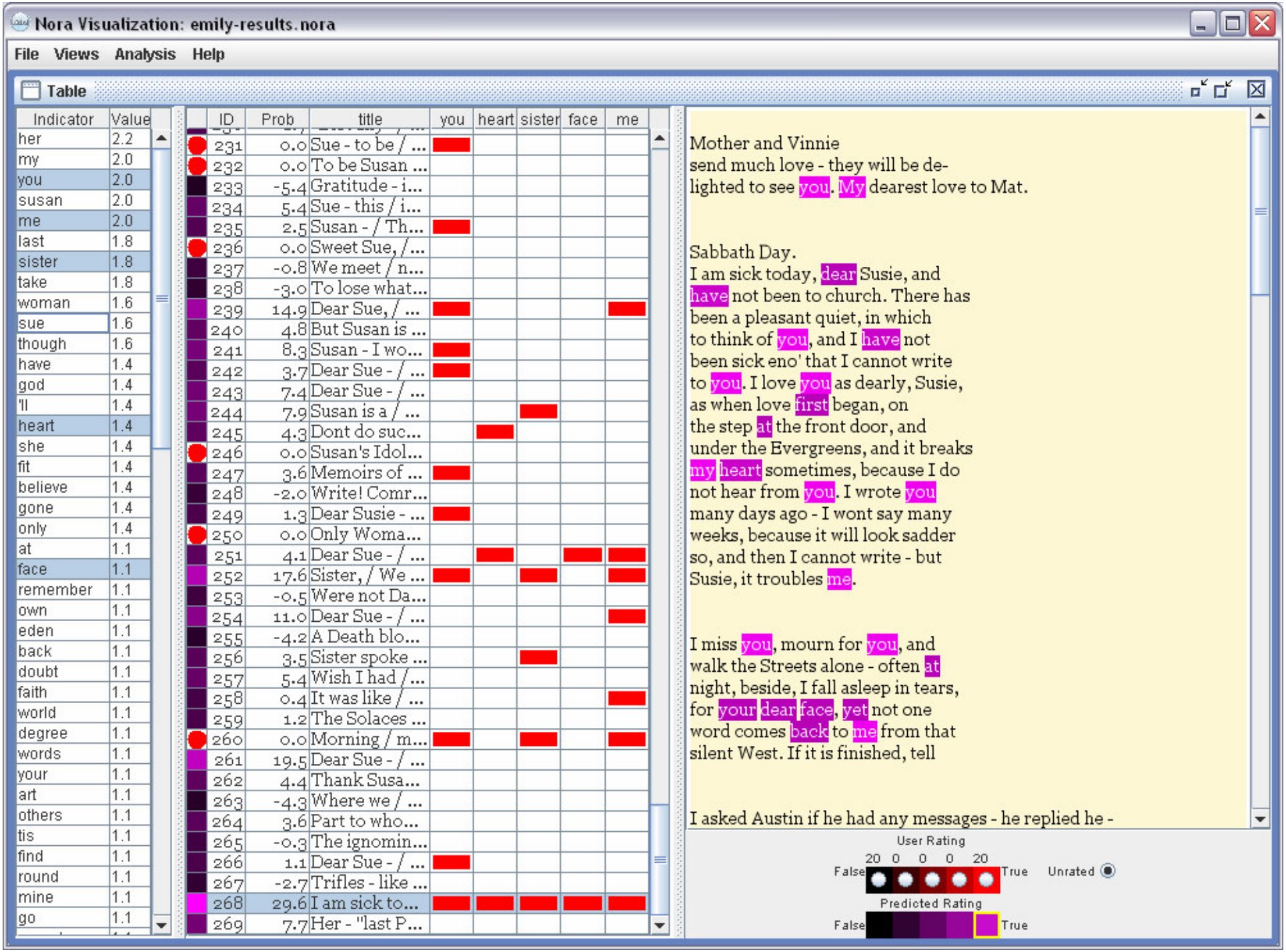
- we use the short word “hot”

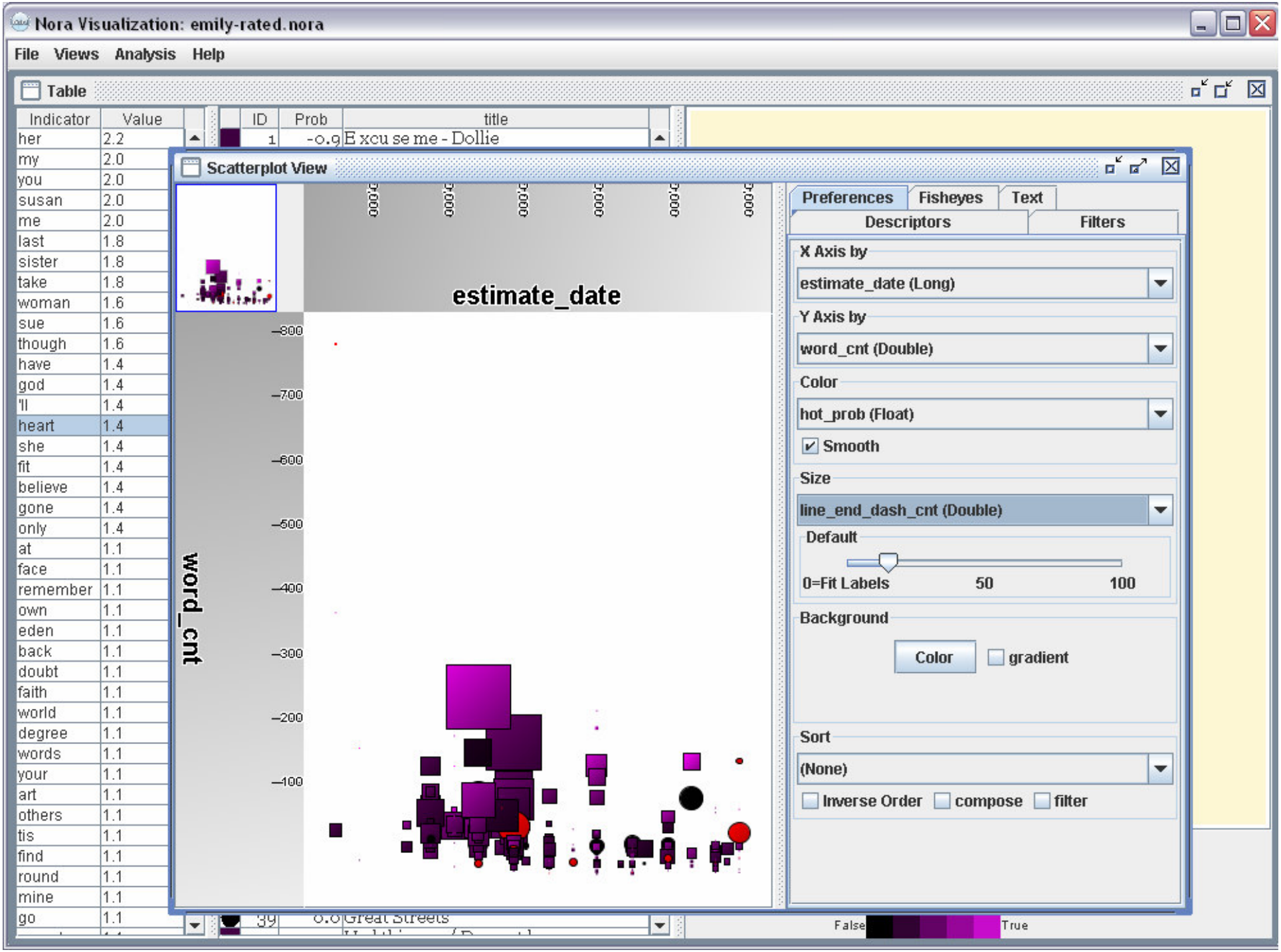

Nora Visualization: emily-rated.nora
File Views Analysis Help
Table
ID title
1 E xcu se me - Dollie
2 Her breast is fit for pearls
3 As Watchers hang upon the East
4 These are the days when Birds come back -
5 Besides the autumn poets sing
6 A slash of Blue - / A sweep of Gray -
7 The Soul unto itself / Is an imperial friend
8 The face I carry with / me - last -
9 The
10 A full fed Rose
11 Ah' Teneriffe! / Retreating Mountain!
12 A Lady red, amid the Hill
13 A little bread, a crust - a crumb
14 A little / Madness in / the Spring
15 All Circumstances are / the Frame
16 All I may - if / small
17 Ambition cannot find him!
18 A prompt - executive / Bird is the Jay -
19 A Sparrow took
20 A Spider sewed / at Night / Without a Light
21 Bloom upon the / Mountain
22 By homely / gifts
23 Content of fading / is enough for me
24 Crisis is sweet and / yet the Heart
25 Delayed till she had ceased to know
26 Defeat - whets Victory - / they say -
27 Dropped into the / Ether Acre!
28 Dust is the only secret -
29 Essential Oils are / wrung -
30 Except the smaller / size
31 Except to Heaven - she is nought.
32 Exhiliration is the Breeze
33 Best Witchcraft is
34 Experiment to Me
35 Exultation is the going / Of an inland soul to
36 Frequently the woods are pink -
37 Given in Marriage / unto Thee
38 Glowing is her Bonnet
39 Great Streets
Her breast is fit for pearls,
But I was not a "Diver" -
Her brow is fit for thrones
But I have not a crest,
Her heart is fit for home -
I - a Sparrow - build there
Sweet of twigs and twine
My perennial nest.
Emily -
Question: Does this document show signs of erotic language?
False (i.e. not hot) True (i.e. hot)
User Rating
20 0 0 0 20
False True Unrated
Predicted Rating
False True


Nora Visualization: emily-rated.nora
File
Views
Analysis
Help
Table
ID
title
1 E xcu se me - Dollie
2 Her breast is fit for pearls
3 As Watchers hang upon the East
4 These are the days when Birds come back -
5 Besides the autumn poets sing
6 A slash of Blue - / A sweep of Gray -
7 The Soul unto itself / Is an imperial friend
8 The face I carry with / me - last -
9 The
10 A full fed Rose
11 Ah' Teneriffe! / Retreating Mountain!
12 A Lady red, amid the Hill
13 A little bread, a crust - a crumb
14 A little / Madness in / the Spring
15 All Circumstances are / the Frame
16 All I may - if / small
17 Ambition cannot find him!
18 A prompt - executive / Bird is the Jay -
19 A Sparrow took
20 A Spider sewed / at Night / Without a Light
21 Bloom upon the / Mountain
22 By homely / gifts
23 Content of fading / is enough for me
24 Crisis is sweet and / yet the Heart
25 Delayed till she had ceased to know
26 Defeat - whets Victory - / they say -
27 Dropped into the / Ether Acre!
28 Dust is the only secret -
29 Essential Oils are / wrung -
30 Except the smaller / size
31 Except to Heaven - she is nought.
32 Exhiliration is the Breeze
33 Best Witchcraft is
34 Experiment to Me
35 Exultation is the going / Of an inland soul to
36 Frequently the woods are pink -
37 Given in Marriage / unto Thee
38 Glowing is her Bonnet
39 Great Streets
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!
User Rating
20 0 0 0 20
False
True
Unrated
Predicted Rating
False
True


Nora Visualization: emily-rated.nora
File
Views
Analysis
Help
Table
Prediction Parameters...
Prediction
Prediction (offline demo)
Get Metadata
Use nora server
Use norma server
ID
title
197 Has All - a / codicil?
198 To take away our / Sue
199 The Leaves like / Women in
200 I send My / Own, two answe
201 Success is counted sweetest
202 Dear Sue - / It is / sweet you
203 Dear Sue - / I'm thinking / o
204 Perception of an / object co
205 My Sue - / Loo and / Fanny
206 A fresh / Morning
207 The Bumble of a Bee - / A W
208 Dear Sue - / With the / Exce
209 The things of / which we wa
210 Mama and / Sister might / li
211 Now I lay / thee down to / S
212 Best Witchcraft / is Geomet
213 Will my great / Sister accep
214 "Egypt - thou / knew'st" -
215 Please Excuse / Santa Claus
216 For largest Woman's / Heart
217 Thank Sue, but / not tonight
218 Susan - I dreamed / of you
219 Lest any doubt / that we are glad
220 "For Brutus, / as you know"
221 A Spell / cannot be / tattered
222 Great Hungers / feed themselves
223 Susan - / Whoever blesses
224 Never mind / dear -
225 To own a / Susan of / my own
226 White as an / Indian Pipe
227
228 "Thank you" / ebbs - between us
229 Dear Sue. / Your - Riches - / taught me -
230 "Lest any" / Hen
Susan - I dreamed
of you, last
night, and send
a Carnation to
indorse it -
Sister of Ophir -
Progress:
fields: document.doc_id,filename,token FROM event,document,token WHERE
fields: document.doc_id,filename,token
0
1
2
prob: -296.5077129191358 ratio: 3.672128384026337 DEAmsEDC
prob: -304.77576030784854 ratio: 8.612149090299795 DEAmsEDC
prob: -868.8382427174977 ratio: 19.501724122962173 DEAmsEDC
prob: -422.4776524897895 ratio: 3.646983932968169 DEAmsEDC
Cancel
User Rating
20 0 0 0 20
False
True
Unrated
Predicted Rating
False
True


Nora Visualization: emily-results.nora
File Views Analysis Help
Table
Indicator Value
her 2.2
my 2.0
you 2.0
susan 2.0
me 2.0
last 1.8
sister 1.8
take 1.8
woman 1.6
sue 1.6
though 1.6
have 1.4
god 1.4
'll 1.4
heart 1.4
she 1.4
fit 1.4
believe 1.4
gone 1.4
only 1.4
at 1.1
face 1.1
remember 1.1
own 1.1
eden 1.1
back 1.1
doubt 1.1
faith 1.1
world 1.1
degree 1.1
words 1.1
your 1.1
art 1.1
others 1.1
tis 1.1
find 1.1
round 1.1
mine 1.1
go 1.1
ID Prob title
231 0.0 Sue - to be / lovely as you
232 0.0 To be Susan / is Imagination
233 -5.4 Gratitude - is not / the mention / Of a
234 5.4 Sue - this / is the last / flower -
235 2.5 Susan - / The sweetest / acts
236 0.0 Sweet Sue, / There is / no first, or last
237 -0.8 We meet / no Stranger / but Ourself.
238 -3.0 To lose what we / never owned
239 14.9 Dear Sue, / God bless you for the bread!
240 4.8 But Susan is / a Stranger yet -
241 8.3 Susan - I would / have come out / of Eden
242 3.7 Dear Sue - / The Supper / was delicate / ...
243 7.4 Dear Sue - / I should love dearly
244 7.9 Susan is a / vast and sweet / Sister
245 4.3 Dont do such / things, dear Sue -
246 0.0 Susan's Idolater keeps / a Shrine
247 3.6 Memoirs of Little / Boys that live -
248 -2.0 Write! Comrade - write! /
249 1.3 Dear Susie - I send / you a little air -
250 0.0 Only Woman / in the World
251 4.1 Dear Sue - / Your litte / mental gallantries
252 17.6 Sister, / We both are / Women
253 -0.5 Were not Day / of itself memo-/rable
254 11.0 Dear Sue - / Just say one / word
255 -4.2 A Death blow - is / a Life blow -
256 3.5 Sister spoke / of Springfield -
257 5.4 Wish I had / something vital
258 0.4 It was like / a breath from / Gibraltar
259 1.2 The Solaces / of Theft
260 0.0 Morning / might come / by Accident
261 19.5 Dear Sue - / A Promise / is firmer
262 4.4 Thank Susan / for the lovely / Supper
263 -4.3 Where we / owe but a / little
264 3.6 Part to whom /
265 -0.3 The ignominy / to receive -
266 1.1 Dear Sue - / You cant think / how much ...
267 -2.7 Trifles - like / Life - and / the Sun,
268 29.6 I am sick today, dear Susie, and / have n...
269 7.7 Her - "last Poems" - / Poets ended -
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 yours, or a
Loaf of mine, spread, you
understand - Great times -
Love for Fanny.
Wish Pope to Rome - thats
all -
Emily.
Esqr. in parlor -
User Rating
20 0 0 0 20
False True Unrated
Predicted Rating
False True


Nora Visualization: emily-results.nora
File Views Analysis Help
Table
Indicator Value
her 2.2
my 2.0
you 2.0
susan 2.0
me 2.0
last 1.8
sister 1.8
take 1.8
woman 1.6
sue 1.6
though 1.6
have 1.4
god 1.4
'll 1.4
heart 1.4
she 1.4
fit 1.4
believe 1.4
gone 1.4
only 1.4
at 1.1
face 1.1
remember 1.1
own 1.1
eden 1.1
back 1.1
doubt 1.1
faith 1.1
world 1.1
degree 1.1
words 1.1
your 1.1
art 1.1
others 1.1
tis 1.1
find 1.1
round 1.1
mine 1.1
go 1.1
ID Prob title you heart sister face me
231 0.0 Sue - to be / ...
232 0.0 To be Susan ...
233 -5.4 Gratitude - i...
234 5.4 Sue - this / i...
235 2.5 Susan - / Th...
236 0.0 Sweet Sue, / ...
237 -0.8 We meet / n...
238 -3.0 To lose what...
239 14.9 Dear Sue, / ...
240 4.8 But Susan is ...
241 8.3 Susan - I wo...
242 3.7 Dear Sue - / ...
243 7.4 Dear Sue - / ...
244 7.9 Susan is a / ...
245 4.3 Dont do suc...
246 0.0 Susan's Idol...
247 3.6 Memoirs of ...
248 -2.0 Write! Comr...
249 1.3 Dear Susie - ...
250 0.0 Only Woma...
251 4.1 Dear Sue - / ...
252 17.6 Sister, / We ...
253 -0.5 Were not Da...
254 11.0 Dear Sue - / ...
255 -4.2 A Death blo...
256 3.5 Sister spoke ...
257 5.4 Wish I had / ...
258 0.4 It was like / ...
259 1.2 The Solaces ...
260 0.0 Morning / m...
261 19.5 Dear Sue - / ...
262 4.4 Thank Susa...
263 -4.3 Where we / ...
264 3.6 Part to who...
265 -0.3 The ignomin...
266 1.1 Dear Sue - / ...
267 -2.7 Trifles - like ...
268 29.6 I am sick to...
269 7.7 Her - "last P...
Mother and Vinnie
send much love - they will be de-
lighted 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 eno' 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 wont 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 -
User Rating
20 0 0 0 20
False True Unrated
Predicted Rating
False True


Nora Visualization: emily-rated.nora
File Views Analysis Help
Table
Indicator Value
her 2.2
my 2.0
you 2.0
susan 2.0
me 2.0
last 1.8
sister 1.8
take 1.8
woman 1.6
sue 1.6
though 1.6
have 1.4
god 1.4
'll 1.4
heart 1.4
she 1.4
fit 1.4
believe 1.4
gone 1.4
only 1.4
at 1.1
face 1.1
remember 1.1
own 1.1
eden 1.1
back 1.1
doubt 1.1
faith 1.1
world 1.1
degree 1.1
words 1.1
your 1.1
art 1.1
others 1.1
tis 1.1
find 1.1
round 1.1
mine 1.1
go 1.1
ID Prob title
1 -0.9 E xcu se me - Dollie
39 0.0 Great Streets
Scatterplot View
estimate_date
word_cnt
800
700
600
500
400
300
200
100
Preferences Fisheyes Text
Descriptors Filters
X Axis by
estimate_date (Long)
Y Axis by
word_cnt (Double)
Color
hot_prob (Float)
Smooth
Size
line_end_dash_cnt (Double)
Default
0=Fit Labels 50 100
Background
Color gradient
Sort
(None)
Inverse Order compose filter
False True

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

HCIL

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

HCIL