# All about me….

Tom Horton

Dept. of Comp. Sci.

Univ. of Virginia

# My Areas of Interest

- CS and Software Engineering Education
- Software Engineering and Development
- Humanities Computing

# Software Engineering and Development

- My interests in these topics are primarily:
  - – How to teach such things better
  - – How to apply them to certain problem areas
- Software Design
  - – Stuff taught in CS4240 (patterns, architecture)
  - – SW architecture (particularly in reusable designs for specific problem domains)
  - – HCI and usability (CS3205)
    - including evaluation experiments

# CS and SW Engin. Education

- Important:  Needs to be an <u>education research</u> oriented project, which means:
  - create something new
  - try it out experimentally
  - evaluate results
- Doing this with real subjects can:
  - require lots of advance planning
  - be time-consuming
  - be messy

# CS and SW Engin. Education

- Possible topic areas:
  - How do beginning students learn how to code, or use tools?
  - How do more experienced students learn how to design, debug, problem-solve?
  - Usability of tools (like IDEs, debuggers, design tools)
  - Creating tools or environments to support education
    - collecting feedback on student learning, etc.

# Humanities Computing Overview

Dr. Tom Horton

Dept. of Computer Science

University of Virginia

horton@virginia.edu

# Overview

- Background: text processing, humanities computing
- Past Explorations
  - A region-based model of documents
  - Text mining and sentimentalism
  - Recent work with Dolley Madison letters

# Humanities Computing: An “Old” User Community

- 1964 Literary Data Processing Conference
  - Papers on corpus preparation, stylistics, dictionaries
  - Most common software tool: concordance program
- Joseph Raben’s research problem:
  - Given two texts, find pairs of sentences that contain verbal echoes
  - Shelley’s *Prometheus Unbound* heavily influenced by the language of Milton’s *Paradise Lost*
  - Raben’s papers emphasize algorithm

# Humanities Computing

- Applying software, algorithms, etc. to problems of interest to humanities scholars
  - In particular, literary text analysis
  - Tools for finding things, features, “chunks”, or showing relationships between texts
- Data mining and text mining
- Information visualization

- Working with real users, real problems here at UVa

# My Interests

- Improving our ability to develop new text processing and analysis software tools for humanities users
  - Based on my viewpoint as a software engineer
  - Includes study of user requirements, designs, frameworks, reusable components
- We could develop a family of software tools that:
  - satisfy common core requirements in the same way
  - share common core concepts and approaches
  - are based on a general model of text processing
  - that use a flexible software architecture

# Background: Text Processing and Humanities Computing

- Users: scholars studying texts, linguists, etc. for the purpose of
  - preparing editions,
  - carrying out stylistic or authorship analyses,
  - finding relationships between multiple texts,
  - finding parts of a text (perhaps in a large corpus) with certain traits,
- Characteristics:
  - Texts often not modern English
  - Mark-up such as XML very important

# An “Old” User Community

- 1964 Literary Data Processing Conference
  - Papers on corpus preparation, stylistics, dictionaries
  - Most common software tool: concordance program
- Joseph Raben’s research problem:
  - Given two texts, find pairs of sentences that contain verbal echoes
  - Shelley’s *Prometheus Unbound* heavily influenced by the language of Milton’s *Paradise Lost*
  - Raben’s papers emphasize algorithm

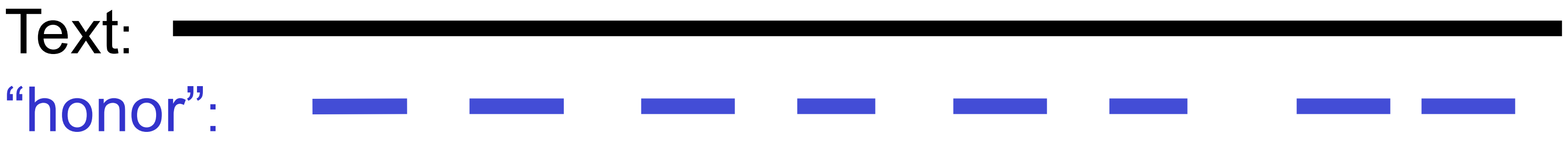
# Software Tools Needed

- Few software tools have been developed
  - The user community recognizes this a major problem
  - Example: No easy to use “app” is out there to solve Raben’s problem
- Reasons:
  - Community is dispersed and is not resource rich
  - Supporting multi-lingual definitions of alphabets, collation sequence, etc. and their output
  - Recently: SGML markup adds complexity
  - Users need good user interfaces

# 1. Regions in Text

# Example: Regions and Region Sets

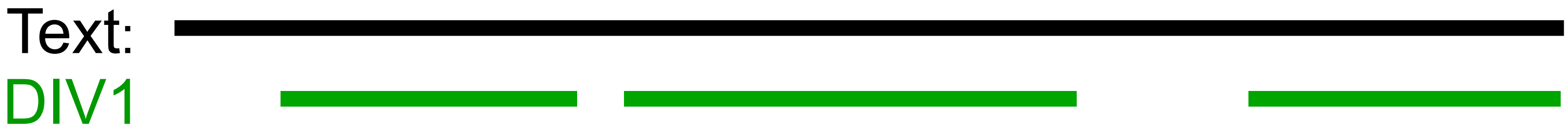
## (1) All Occurrences of “honor”:



Text:

“honor”:

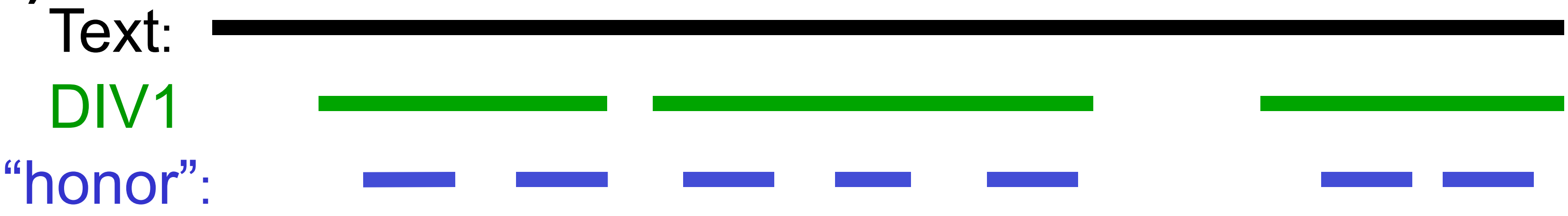
## (2) All Occurrences of DIV1 elements:



Text:

DIV1

## (3) All Occurrences of “honor” in a DIV1 element:



Text:

DIV1

“honor”:

# Region Examples

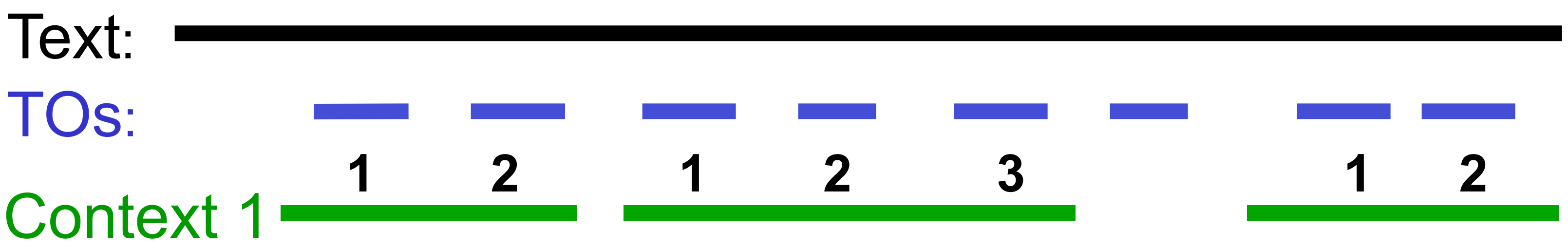
- Region sets:
  - all words in a text; all characters; all syllables
  - all occurrences of a given token
  - all DIV1 elements
  - all elements that have attribute with a given value
- Queries or subtext selection. Examples: Let’s find:
  - Speeches by Hamlet in Acts 4 and 5,...
  - choosing only those marked-up as verse,...
  - choosing only those with the word “honor”
- This example illustrates *selection of a* document or subdocument, a core user requirement

# Benefits of Using Regions

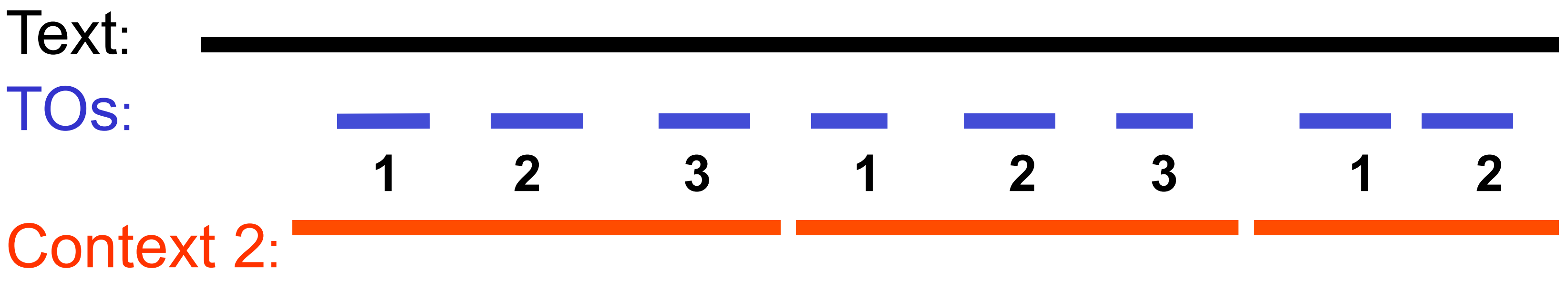
- Provides a **general** model of things in texts
  - Markup such as XML elements can be modeled as regions
  - sgrep’s model of regions uses a concept of nesting or inclusion, which is naturally useful
- A good model can be quite powerful
  - Spreadsheets: cells, rows, columns, ranges

# Example: Sequencing TOs

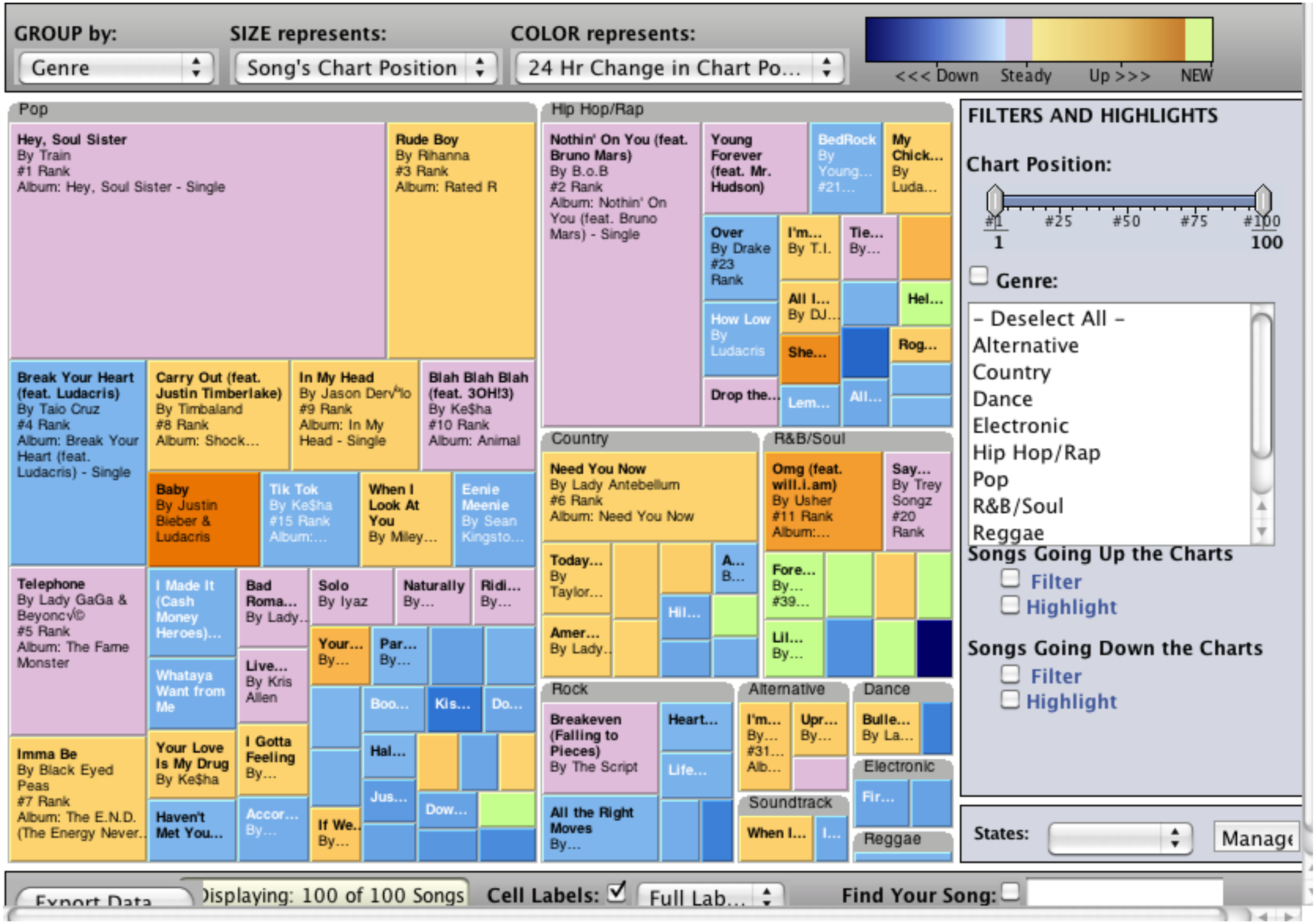
## Numbered TOs in Some Context, Ignoring Some:



## Numbered TOs in Another Context:



# 2. Visualization



GROUP by: Genre

SIZE represents: Song's Chart Position

COLOR represents: 24 Hr Change in Chart Po…

<<< Down Steady Up >>> NEW

FILTERS AND HIGHLIGHTS

Chart Position: 1 – 100

□ Genre:

- Deselect All -
Alternative
Country
Dance
Electronic
Hip Hop/Rap
Pop
R&B/Soul
Reggae

Songs Going Up the Charts
□ Filter
□ Highlight

Songs Going Down the Charts
□ Filter
□ Highlight

States:

Manage

Export Data

Displaying: 100 of 100 Songs

Cell Labels: ☑ Full Lab…

Find Your Song: □

#1
#25
#50
#75
#100
Pop
Hey, Soul Sister By Train #1 Rank Album: Hey, Soul Sister - Single
Rude Boy By Rihanna #3 Rank Album: Rated R
Break Your Heart (feat. Ludacris) By Taio Cruz #4 Rank Album: Break Your Heart (feat. Ludacris) - Single
Carry Out (feat. Justin Timberlake) By Timbaland #8 Rank Album: Shock…
In My Head By Jason Derv^lo #9 Rank Album: In My Head - Single
Blah Blah Blah (feat. 3OH!3) By Ke$ha #10 Rank Album: Animal
Baby By Justin Bieber & Ludacris
Tik Tok By Ke$ha #15 Rank Album:…
When I Look At You By Miley…
Eenie Meenie By Sean Kingsto…
Telephone By Lady GaGa & Beyoncv© #5 Rank Album: The Fame Monster
I Made It (Cash Money Heroes)…
Bad Roma… By Lady…
Solo By Iyaz
Naturally By…
Ridi… By…
Whataya Want from Me
Live… By Kris Allen
Your… By…
Par… By…
Boo…
Kis…
Do…
Imma Be By Black Eyed Peas #7 Rank Album: The E.N.D. (The Energy Never…
Your Love Is My Drug By Ke$ha
I Gotta Feeling By…
Hal…
Jus…
Dow…
Haven't Met You…
Accor… By…
If We… By…
Hip Hop/Rap
Nothin' On You (feat. Bruno Mars) By B.o.B #2 Rank Album: Nothin' On You (feat. Bruno Mars) - Single
Young Forever (feat. Mr. Hudson)
BedRock By Young… #21…
My Chick… By Luda…
Over By Drake #23 Rank
I'm… By T.I.
Tie… By…
All I… By DJ…
Hel…
How Low By Ludacris
She…
Rog…
Drop the…
Lem…
All…
Country
Need You Now By Lady Antebellum #6 Rank Album: Need You Now
Today… By Taylor…
A… B…
Hil…
Amer… By Lady…
R&B/Soul
Omg (feat. will.i.am) By Usher #11 Rank Album:…
Say… By Trey Songz #20 Rank
Fore… By… #39…
Lil… By…
Rock
Breakeven (Falling to Pieces) By The Script
Heart…
Life…
All the Right Moves By…
Alternative
I'm… By… #31… Alb…
Upr… By…
Soundtrack
When I…
I…
Dance
Bulle… By La…
Electronic
Fir…
Reggae

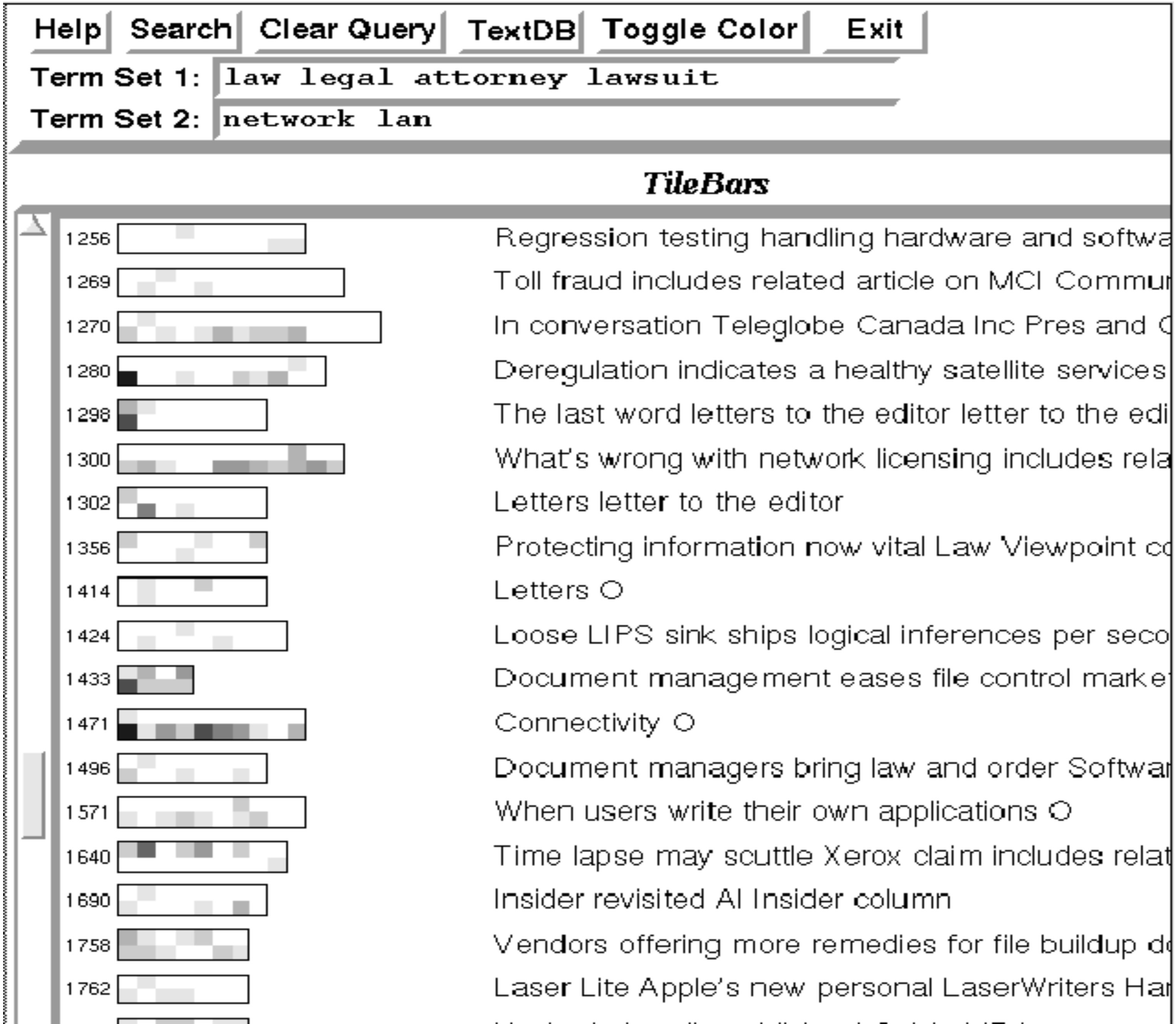
# Visualization

- Earlier slide shows a simple visualization of TOs inside of regions
- Potential for a general model of visualization of TOs (words, markup elements etc.) in relation to other TOs (regions, markup elements, etc.)
- Examples:
  - Show presence or absence of some TO in a selected region
  - Counts of how many occurrences in each region
- Given a region-based database of TO Occurrences, we have a **dynamic** and **flexible** way to visualize all TOs known to the tool

# Tilebars

- Marti Hearst developed tilebars as a method for visualizing query results
- Dimensions:
  - documents,
  - terms (text objects),
  - segments (regions)
- Shading of each cell shows *strength of occurrence* of a term
- Adjacent shaded cells show *co-occurrence* of terms

# Tilebars (example)



Term Set 1: law legal attorney lawsuit

Term Set 2: network lan

| | |
|---|---|
| 1256 | Regression testing handling hardware and softwa |
| 1269 | Toll fraud includes related article on MCI Commu |
| 1270 | In conversation Teleglobe Canada Inc Pres and C |
| 1280 | Deregulation indicates a healthy satellite services |
| 1298 | The last word letters to the editor letter to the edi |
| 1300 | What's wrong with network licensing includes rela |
| 1302 | Letters letter to the editor |
| 1356 | Protecting information now vital Law Viewpoint co |
| 1414 | Letters ○ |
| 1424 | Loose LIPS sink ships logical inferences per seco |
| 1433 | Document management eases file control marke |
| 1471 | Connectivity ○ |
| 1496 | Document managers bring law and order Softwa |
| 1571 | When users write their own applications ○ |
| 1640 | Time lapse may scuttle Xerox claim includes relat |
| 1690 | Insider revisited AI Insider column |
| 1758 | Vendors offering more remedies for file buildup d |
| 1762 | Laser Lite Apple's new personal LaserWriters Ha |

# Bird Flu Sample Text

On 15 July 2005, the French authorities were informed by the British authorities of a confirmed outbreak of Newcastle disease (ND) on a pheasant farm in Surrey, England. The French authorities immediately launched an epidemiological investigation to determine whether French farms could be at risk.

The results of this investigation revealed that five farms located in two French departements - one in Loire-Atlantique and four in Vendie - had supplied the affected English farm with pheasants in three consignments between 21 June 2005 and 5 July 2005. The five farms were immediately blocked and placed under surveillance.

Also on 15 July 2005, veterinary inspections as well as serological and virological sampling were conducted on these farms.

# Bird Flu Text Tagged

```
0000002 010  II      On                       Z5
0000002 020  MC      15                       N1 T1.2 T3 T1.3 N3.2
0000002 030  NPM1    July                     T1.3[i1.2.1 T1.3
0000002 040  MC      2005                     T1.3[i1.2.2 N1 T1.2 T3 T1.3 N3.2
0000002 041  ,       ,
0000002 050  AT      the                      Z5
0000002 060  JJ      French                   Z2 Z2/Q3 Z2/S2mfnc S3.2/B1%
0000002 070  NN2     authorities              G1.1 S7.1+ S7.4+ X2.2+
0000002 080  VBDR    were                     Z5 A3+
0000002 090  VVN     informed                 X2.2+ X2.4 Q2.1
0000002 100  II      by                       Z5
0000002 110  AT      the                      Z5
0000002 120  JJ      British                  Z2 Z2/S2mfnc
0000002 130  NN2     authorities              G1.1 S7.1+ S7.4+ X2.2+
0000002 140  IO      of                       Z5
0000002 150  AT1     a                        Z5
0000003 010  JJ@     confirmed                A7+ S9@
0000003 020  NN1     outbreak                 B2-
0000003 030  IO      of                       Z5
0000003 040  NP1     Newcastle                B2-[i2.2.1 Z2
0000003 050  NN1     disease                  B2-[i2.2.2 B2-
0000003 051  (       (
0000003 060  NP1     ND                       Z99
```

# 3. Recent Dolley Madison Work

- Joe Berger
  - 4th year SEAS CS major
- Papers, letters by Founding “Persons”
  - Find, group, display similar documents
  - Annotate different documents
    - Process annotations
  - Etc.

# 4. Text Mining

- Text-mining for literary research
  - Example: Sentimentalism
  - Other examples:
    - Eroticism in Emily Dickinson
    - Vocabulary in papers on literary criticism

# Example: nora's Sentimentalism Study

- Apply nora ideas to a set of 19th century novels in the Early American Fiction digital library
- Help scholars better understand sentimentalism in a core set of highly sentimental novels
- Identify seemingly sentimental parts of other documents
  - help prove the usefulness of TM in literary criticism

# What is Sentimentalism?

- Term “sentimental novel” first applied to 18th century texts
  - Feeling is valued over reason
  - Author attempts to induce a specific response from the reader
    - Often for a cause: anti-slavery, female education, temperance, etc.
  - Conventional plot devices, characters, repetitions
  - Explicit authorial interventions

# Why It’s an Interesting Problem

- Some novels were hugely popular in the US
- Many novels written by women
- Social issues: e.g. slavery
- Solidification of novel form, and predecessor to Victorian period
- Often used as a derogatory term
  - both then and now
  - but increased recent interest

# Text-Mining for Such Problems

- Data-mining on documents
  - So far: Data (“features”) are vocabulary-based
  - Our first analyses do not use POS, parsing, etc.
- Possible goals:
  - Classification: From a small set of “known” results, make predictions about “unknown” results
    - Explanation?
  - Clustering: Group or organize unknown results based on non-obvious similarities

# Text-Mining Outputs

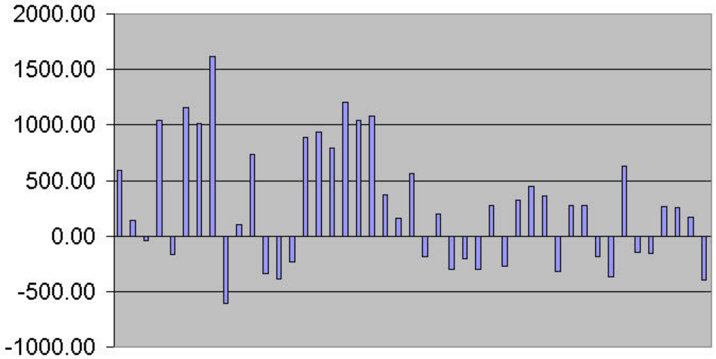
1. A numeric score indicating the degree that a chapter seems sentimental (or not)
   - What’s most sentimental? Least? What’s the pattern?
2. Predictors: vocabulary ordered to show which words contribute most or least to assigning each chapter
   - Possibly a form of explanation for the scholar

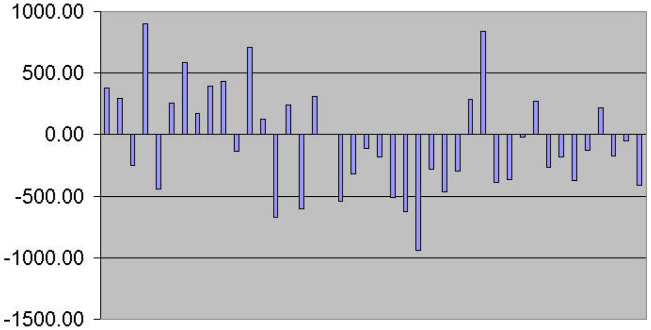
# Sentimental Experiment Plan

- Experiment 1:
  - Goal: To evaluate the use of text-mining on a small set of "core" sentimental novels.
  - Scholars assign a score or label for each chapter in five novels
  - Run text-mining and see what we learn about the methods and the novels

# Change During a Novel

- Stowe’s two novels show more by-chapter variation than Rowson’s works
  - *UTC* has fluctuation between highly-sentimental episodes with scenes of minstrelsy or humor
  - *The Minister’s Wooing* shares this flow (though about marriage)
- Reminder: negative means more sentimental





# Vocabulary Predictors

- The text-mining method used ranks words by how strongly they indicate sentimental or not-sentimental
- Highly-sentimental words include proper names
  - Makes sense: particular characters appear in highly sentimental chapters
  - Won’t lead to models that generalize well for new novels
  - A solution: use part-of-speech tagging to ignore proper-nouns for TM

# Thanks! Discussion?