CS448B :: 17 Nov 2011

Text Visualization

Jason Chuang Stanford University

Why visualize text?

Why visualize text?

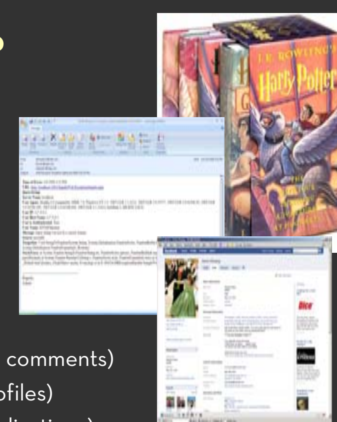
Understanding – get the "gist" of a document

Grouping – cluster for overview or classification

Compare – compare document collections, or inspect evolution of collection over time

Correlate – compare patterns in text to those in other data, e.g., correlate with social network

What is text data?



Documents

- Articles, books and novels
- E-mails, web pages, blogs
- Tags, comments
- Computer programs, logs

Collection of documents

- Messages (e-mail, blogs, tags, comments)
- Social networks (personal profiles)
- Academic collaborations (publications)

Example: Health Care Reform

- Recent history
  - Initiatives by President Clinton
  - Overhaul by President Obama
- Text data
  - News articles
  - Speech transcriptions
  - Legal documents

- **What questions might you want to answer?**
- **What visualizations might help?**

A Concrete Example


September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of yo

Tag Clouds: Word Count

President Obama's Health Care Speech to Congress [New York Times]

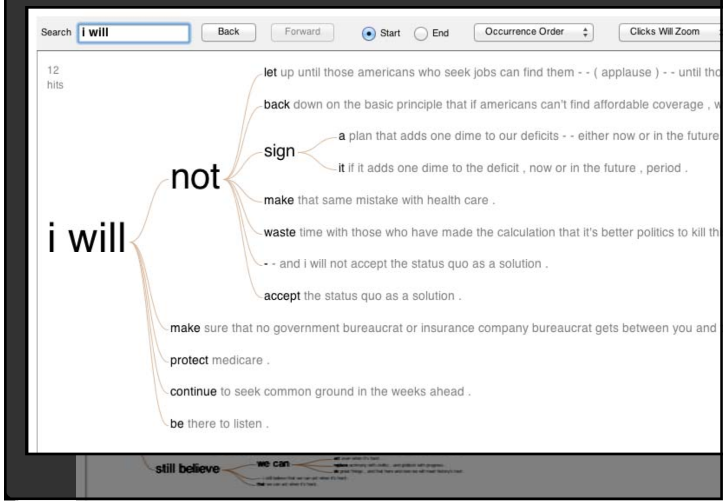
economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

Bill Clinton 1993

Barack Obama 2009

economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

WordTree: Word Sequences



A Double Gulf of Evaluation

Many (most?) text visualizations do not represent the text directly. They represent the output of a language model (word counts, word sequences, etc.).

- Can you **interpret** the visualization? How well does it convey the properties of the model?
- Do you **trust** the model? How does the model enable us to reason about the text?

Challenges of Text Visualization

- High Dimensionality
  - Where possible use **text to represent text**…
    … which terms are the most descriptive?
- Context & Semantics
  - Provide **relevant context** to aid understanding.
  - Show (or provide access to) the **source text**.
- Modeling Abstraction
  - Determine your **analysis task**.
  - Understand abstraction of your **language models**.
  - Match analysis task with appropriate tools and models.

Topics

Text as Data
Visualizing Document Content
Evolving Documents
Visualizing Conversation
Document Collections

Text as Data

Words are (not) nominal?

High dimensional (10,000+)
More than equality tests
Words have meanings and relations

- Correlations: *Hong Kong, San Francisco, Bay Area*
- Order: *April, February, January, June, March, May*
- Membership: *Tennis, Running, Swimming, Hiking, Piano*
- Hierarchy, antonyms & synonyms, entities, …

Text Processing Pipeline

1. Tokenization
   - Segment text into terms.
   - Remove stop words? *a, an, the, of, to, be*
   - Numbers and symbols? *#gocard, @stanfordfball, Beat Cal!!!!!!!!*
   - Entities? *San Francisco, O'Connor, U.S.A.*
2. Stemming
   - Group together different forms of a word.
   - Porter stemmer? *visualization(s), visualize(s), visually* → *visual*
   - Lemmatization? *goes, went, gone* → *go*
3. Ordered list of terms

Tips: Tokenization and Stemming

- Well-formed text to support stemming?
  txt u l8r!
- Word meaning or entities?
  #berkeley → *#berkelei*
- Reverse stems for presentation.
  Ha appl made programm cool?
  Has Apple made programmers cool?

Bag of Words Model

Ignore ordering relationships within the text

A document ≈ vector of term weights

- Each dimension corresponds to a term (10,000+)
- Each value represents the relevance
  - For example, simple term counts

Aggregate into a document-term matrix

- Document vector space model

Document-Term Matrix

Each document is a vector of term weights

Simplest weighting is to just count occurrences

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

WordCount (Harris 2004)

http://wordcount.org

Tag Clouds

- Strength
  - Can help with initial query formation.
- Weaknesses
  - Sub-optimal visual encoding (size vs. position)
  - Inaccurate size encoding (long words are bigger)
  - May not facilitate comparison (unstable layout)
  - Term frequency may not be meaningful
  - Does not show the structure of the text

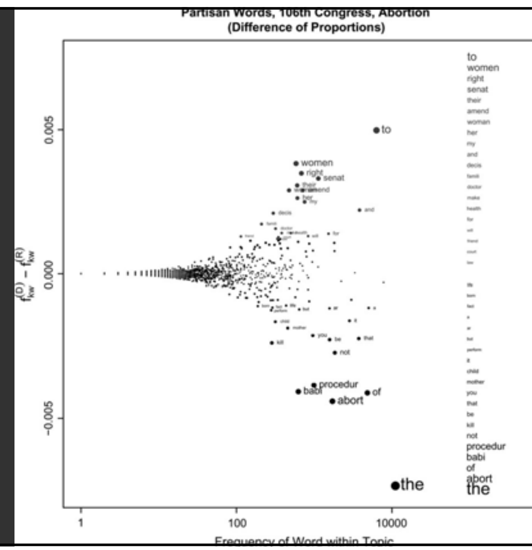
Keyword Weighting

Term Frequency

tf_{td} = count(t) in d

Can take log frequency: $\log(1 + tf_{td})$

Can normalize to show proportion: $tf_{td} / \Sigma_t tf_{td}$



Keyword Weighting

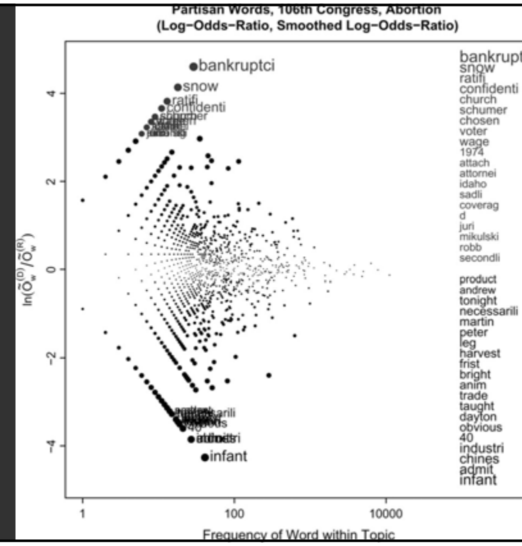
Term Frequency

tf_{td} = count(t) in d

TF.IDF: Term Freq by Inverse Document Freq

$tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$

df_t = # docs containing t; N = # of docs



Keyword Weighting

Term Frequency

tf_{td} = count(t) in d

TF.IDF: Term Freq by Inverse Document Freq

$tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$

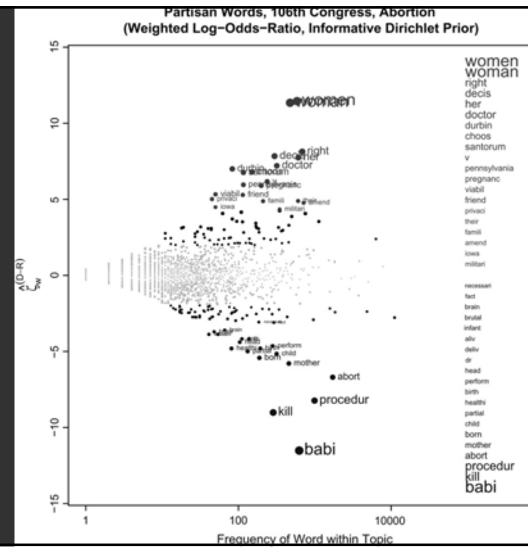
df_t = # docs containing t; N = # of docs

G²: Probability of different word frequency

$E_1 = |d| \times (tf_{td} + tf_{t(C-d)}) / |C|$

$E_2 = |C-d| \times (tf_{td} + tf_{t(C-d)}) / |C|$

$G^2 = 2 \times (tf_{td} \log(tf_{td}/E_1) + tf_{t(C-d)} \log(tf_{t(C-d)}/E_2))$



Limitations of Frequency Statistics?

Typically focus on unigrams (single terms)

Often favors frequent (TF) or rare (IDF) terms

- Not clear that these provide best description

A “bag of words” ignores additional information

- Grammar / part-of-speech
- Position within document
- Recognizable entities

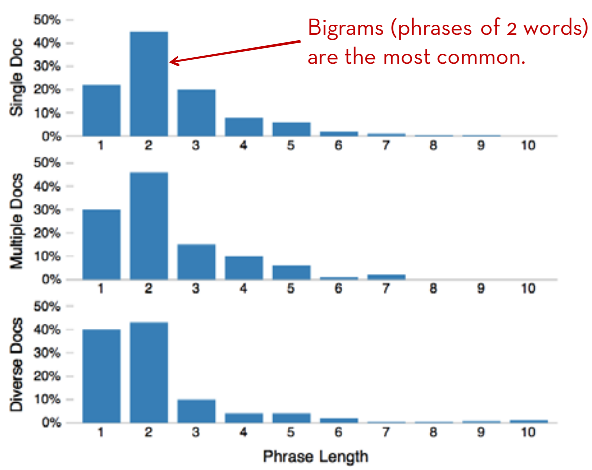
How do people describe text?

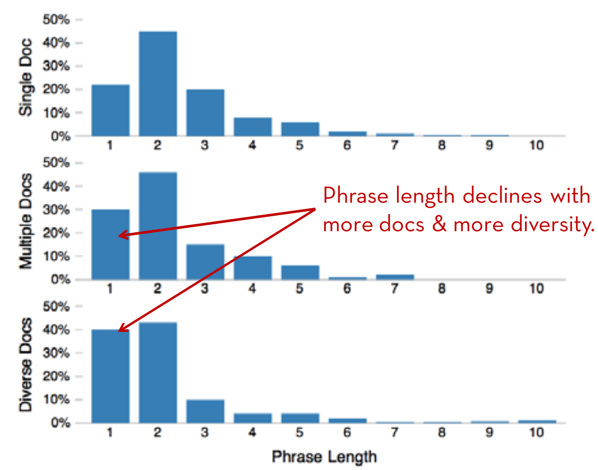
We asked 69 subjects (graduate students) to read and describe dissertation abstracts.

Students were given 3 documents in sequence; they then described the collection as a whole.

Students were matched to both *familiar* and *unfamiliar* topics; *topical diversity* within a collection was varied systematically.

[Chuang, Heer & Manning, 2010]





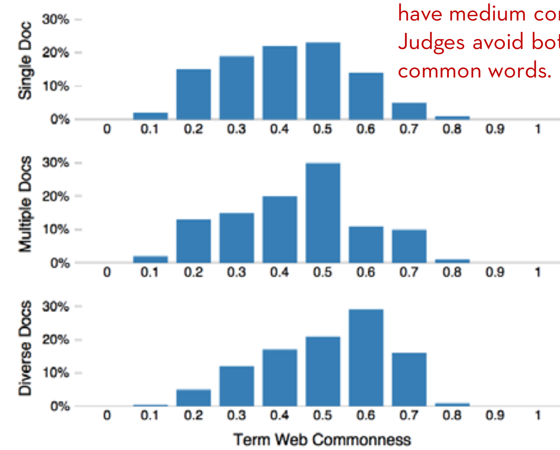
Term Commonness

$$\log(tf_w) \ / \ \log(tf_{the})$$

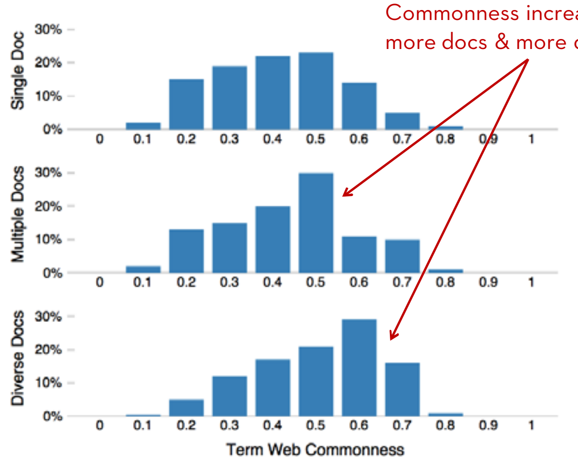
The normalized term frequency relative to the most frequent n-gram, e.g., the word "the".

Measured across an entire corpus or across the entire English language (using Google n-grams)

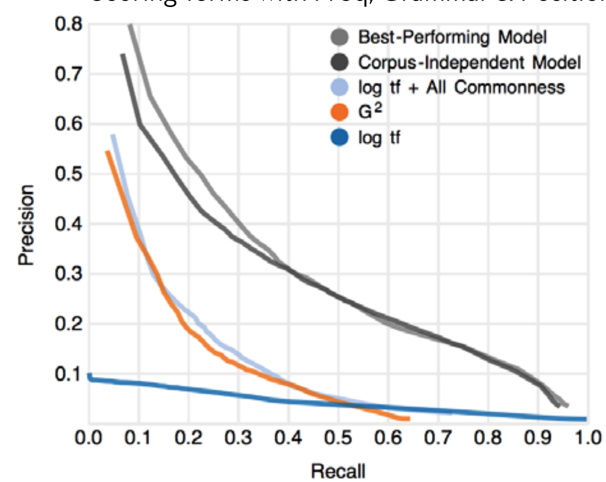
Selected descriptive terms have medium commonness. Judges avoid both rare and common words.

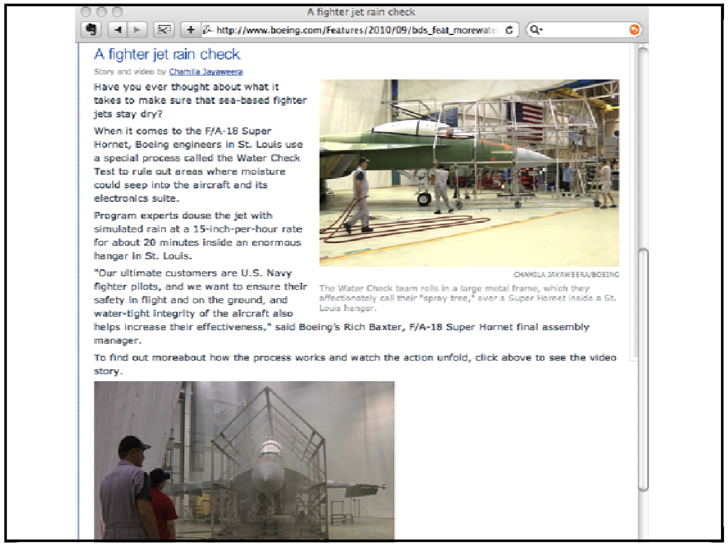


Commonness increases with more docs & more diversity.



Scoring Terms with Freq, Grammar & Position





A fighter jet rain check

A fighter jet rain check

Story and video by Chamila Jayaweera

Have you ever thought about what it takes to make sure that sea-based fighter jets stay dry?

When it comes to the F/A-18 Super Hornet, Boeing engineers in St. Louis use a special process called the Water Check Test to rule out areas where moisture could seep into the aircraft and its electronics suite.

Program experts douse the jet with simulated rain at a 15-inch-per-hour rate for about 20 minutes inside an enormous hangar in St. Louis.

"Our ultimate customers are U.S. Navy fighter pilots, and we want to ensure their safety in flight and on the ground, and water-tight integrity of the aircraft also helps increase their effectiveness," said Boeing's Rich Baxter, F/A-18 Super Hornet final assembly manager.

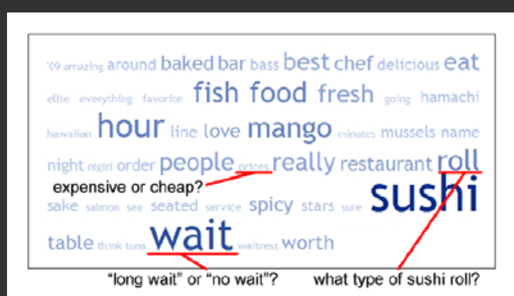
To find out moreabout how the process works and watch the action unfold, click above to see the video story.

CHAMILA JAYAWEERA/BOEING

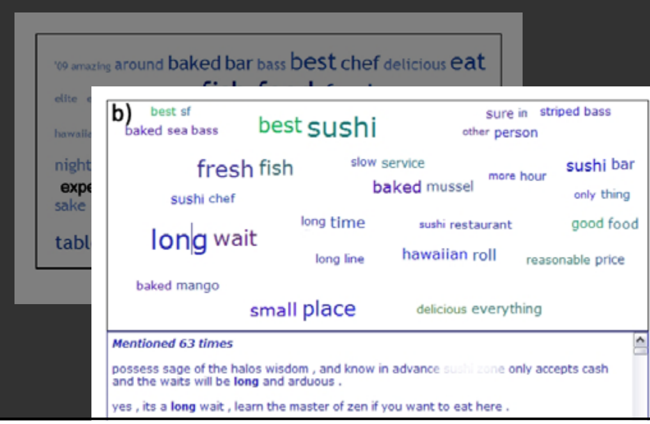
The Water Check team rolls in a large metal frame, which they affectionately call their "spray tree," over a Super Hornet inside a St. Louis hangar.

G^2	*Regression Model*
fighter	Super Hornet
F/A	F/A -18
Hornet	fighter jet
Super	Boeing engineers
Boeing	special process
-18	rain check
rain	electronics suite
St.	Program experts
jet	simulated rain
Louis	ultimate customers
15-inch-per-hour	enormous hangar
douse	water-tight integrity
hangar	Rich Baxter
water-tight	15-inch-per-hour rate
Check	video story
Baxter	aircraft
sea-based	U.S. Navy fighter pilots
aircraft	Super Hornet final assembly manager
Rich	
seep	
click	
Navy	
sure	
Water	
moisture	
watch	
enormous	
stay	

Yelp: Review Spotlight [Yatani 2011]



Yelp: Review Spotlight [Yatani 2011]



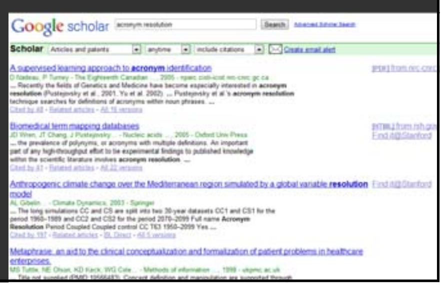
Tips: Descriptive Keyphrases

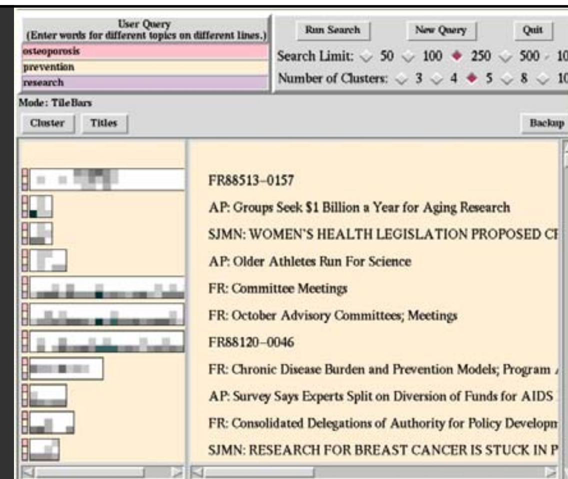
- Understand the limitations of your language model.
  - Bag of words
    - Easy to compute
    - Single words
    - Loss of word ordering
- Select appropriate model and visualization
  - Generate longer, more meaningful phrases
  - Adjective-noun word pairs for reviews
  - Show keyphrases within source text

Visualizing Document Content

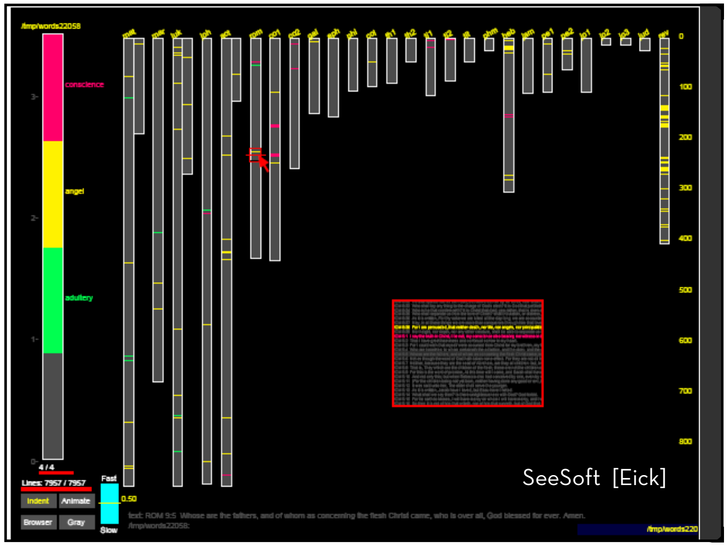
Information Retrieval

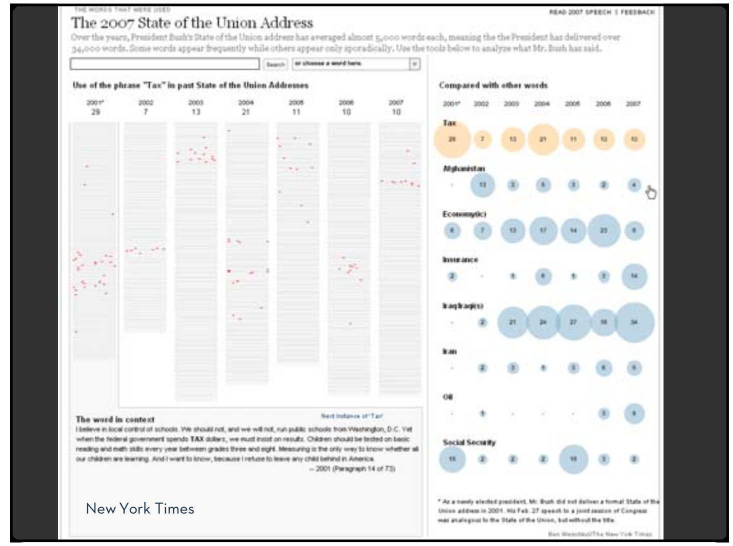
- Search for documents
  - Match query string with documents
- Contextualized search

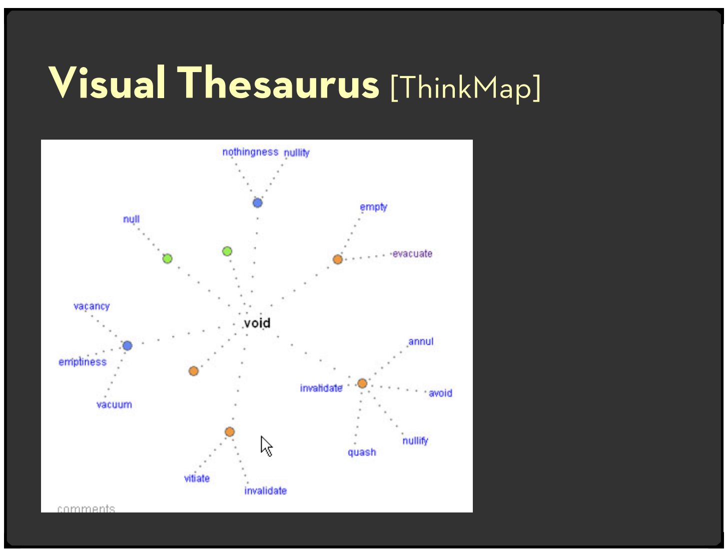




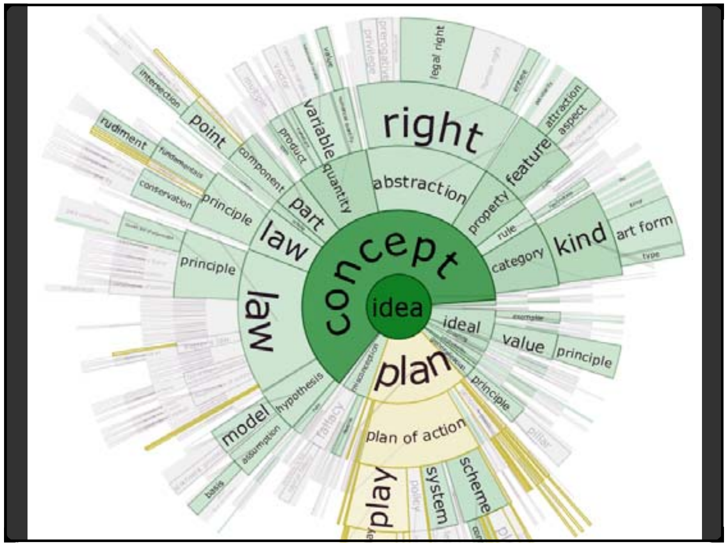
TileBars [Hearst]


SeeSoft [Eick]
conscience
angel
adultery


The 2007 State of the Union Address
Use of the phrase "Tax" in past State of the Union Addresses
Compared with other words
The word in context
New York Times

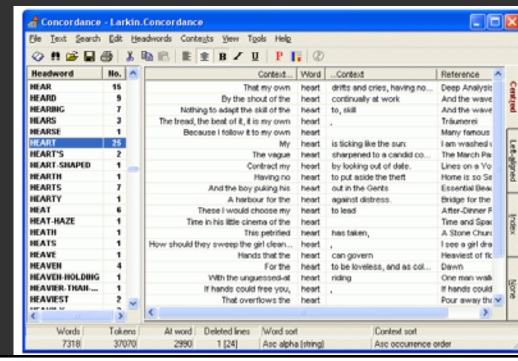

Visual Thesaurus [ThinkMap]
nothingness
nullity
null
empty
evacuate
vacancy
void
emptiness
annul
vacuum
invalidate
avoid
nullify
quash
vitiate
invalidate

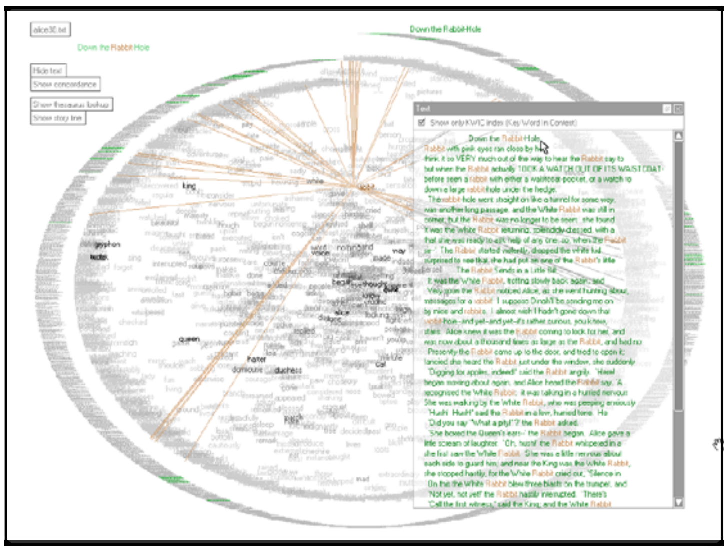

Google
danse libre
Search
Everything
Images
Maps
Videos
News
Shopping
More
The Academy of Danse Libre
About Us
M is for Mazurka
Performances
Danseurs
Auditions
Repertoire
Friday Night Waltz
Elana Waltz (Academy of Danse Libre, 2006-12-08) - YouTube
Danse Libre - Brahms Mazurka Sextille at Stanford Viennese Ball ...

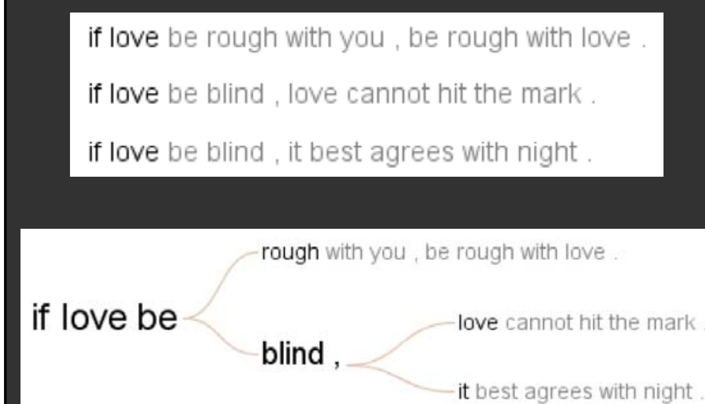


Concordance

What is the common local context of a term?

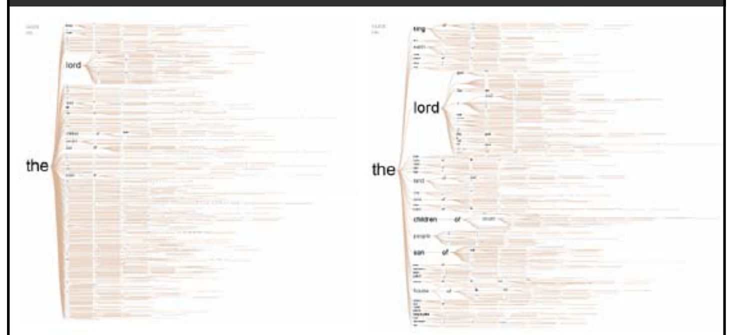




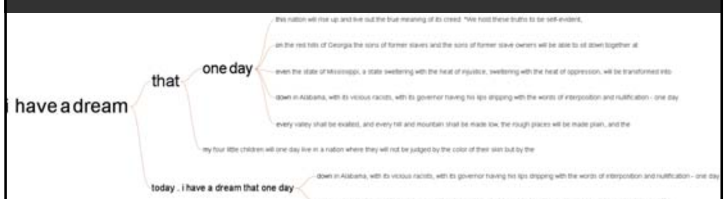


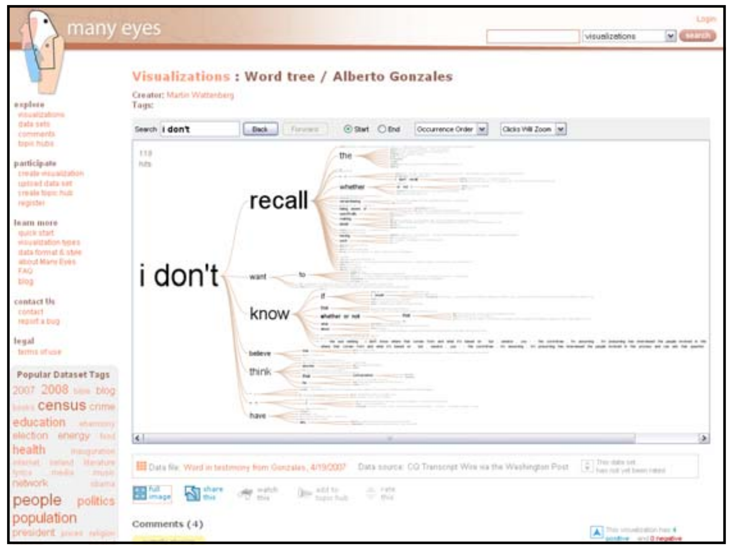
WordTree (Wattenberg et al)

Filter infrequent runs



Recurrent themes in speech





Glimpses of structure

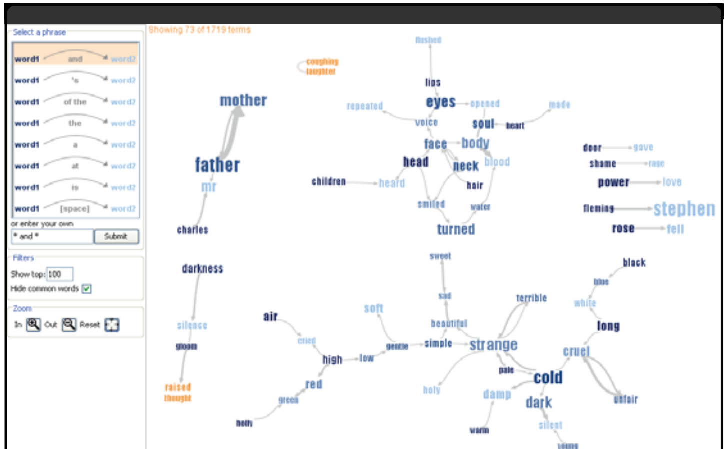
Concordances show local, repeated structure
But what about other types of patterns?

For example
- Lexical: <A> at <B>
- Syntactic: <Noun> <Verb> <Object>

Phrase Nets [van Ham et al]

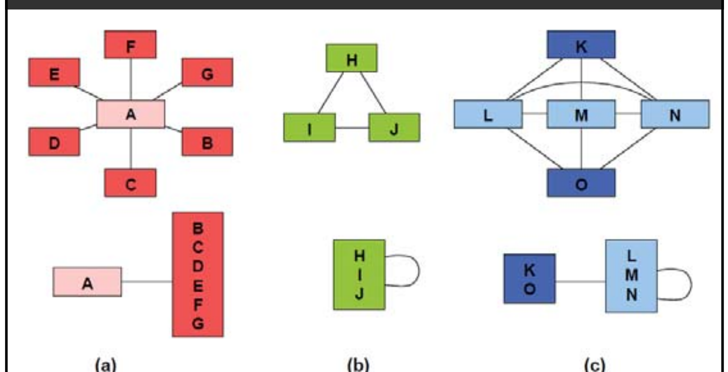
Look for specific linking patterns in the text:
- ‘A and B’, ‘A at B’, ‘A of B’, etc
- Could be output of regexp or parser.

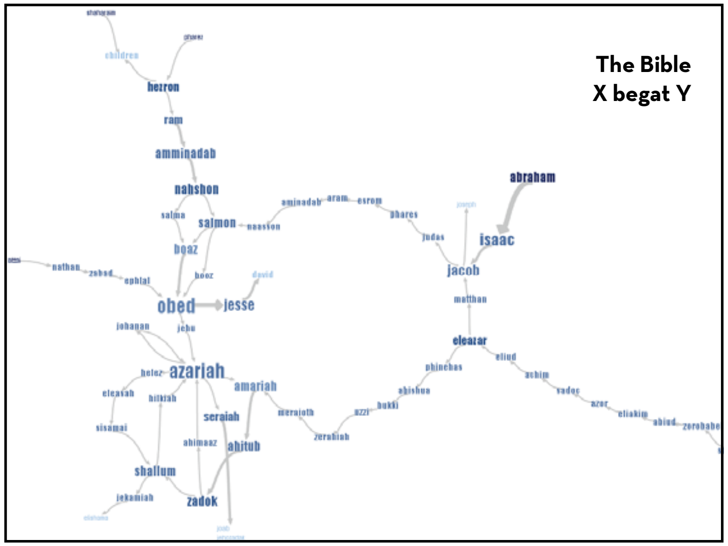
Visualize extracted patterns in a node-link view
- Occurrences → Node size
- Pattern position → Edge direction

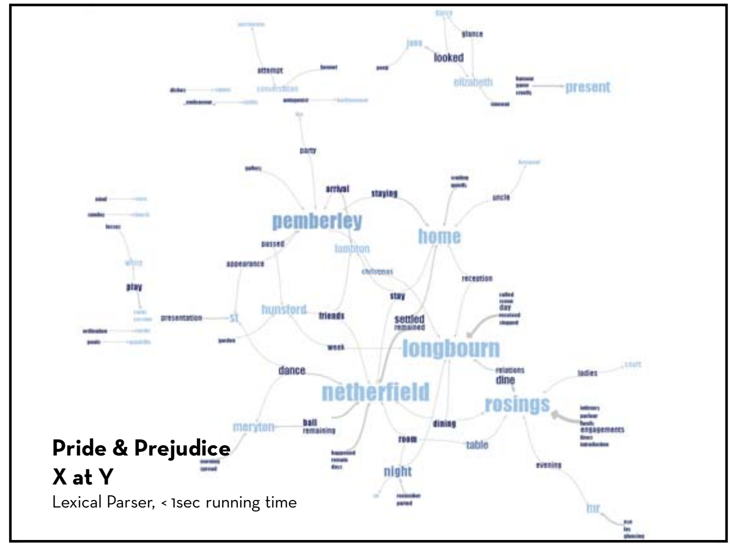


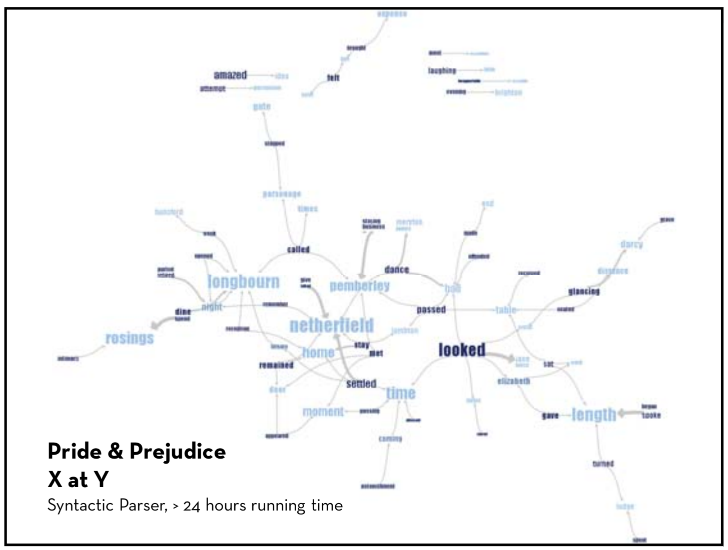
Portrait of the Artist as a Young Man
X and Y

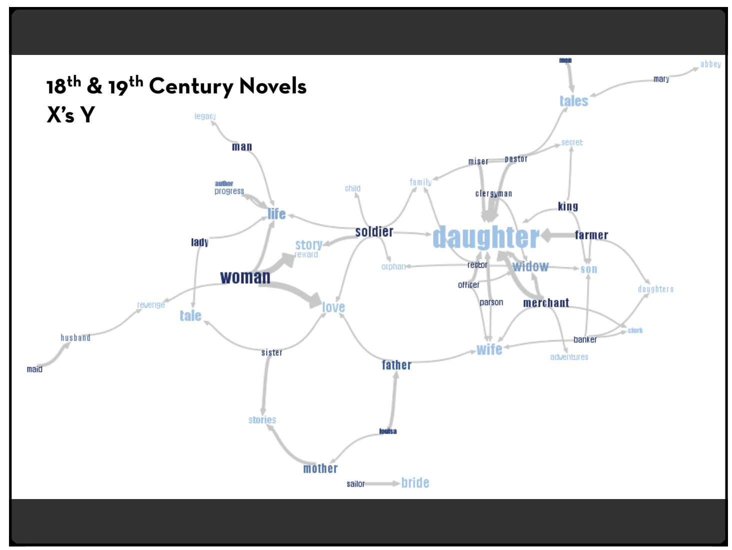
Node Grouping

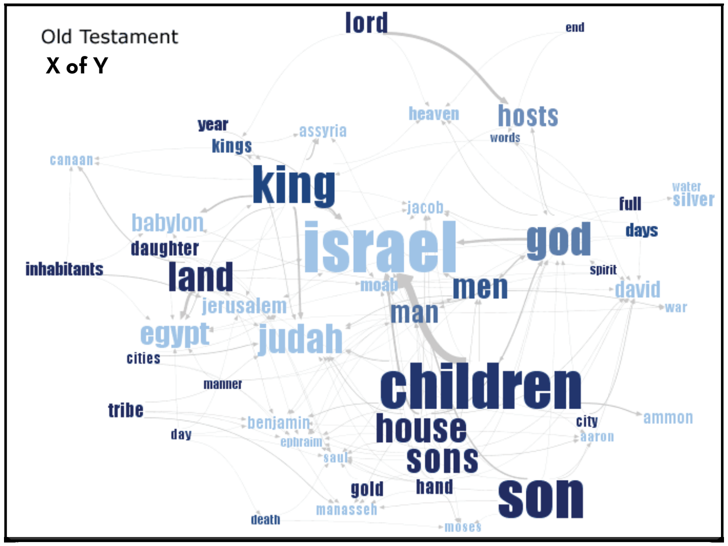


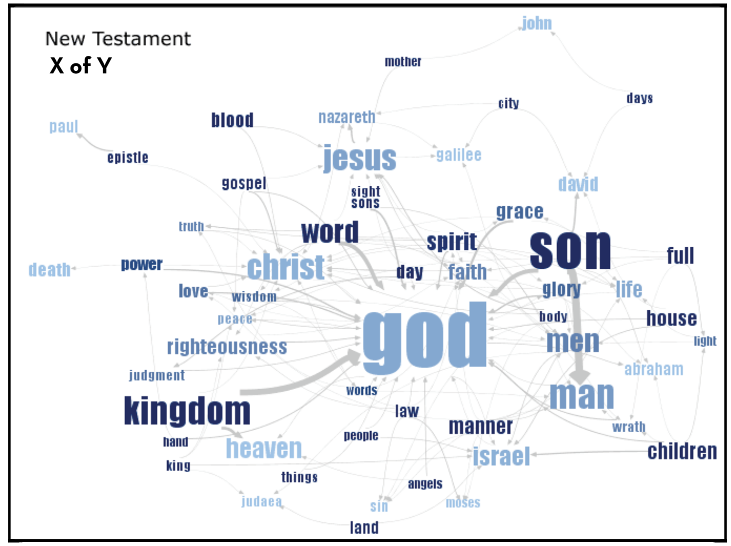

The Bible
X begat Y


Pride & Prejudice
X at Y
Lexical Parser, < 1sec running time


Pride & Prejudice
X at Y
Syntactic Parser, > 24 hours running time


18th & 19th Century Novels
X's Y





Tips: Document Contents

- Understand your task, and handle high dimensionality accordingly…
  - Visually: Word position, browsing, brushing+linking
  - Semantically: Word sequence, hierarchy, clustering
  - Both: Spatial layout reflect semantic relationships
- Role of Interaction:
  - Sufficient language model to enable visual analysis cycles
  - Allow modifications to the model: custom patterns for expressing contextual or domain knowledge

Administrivia

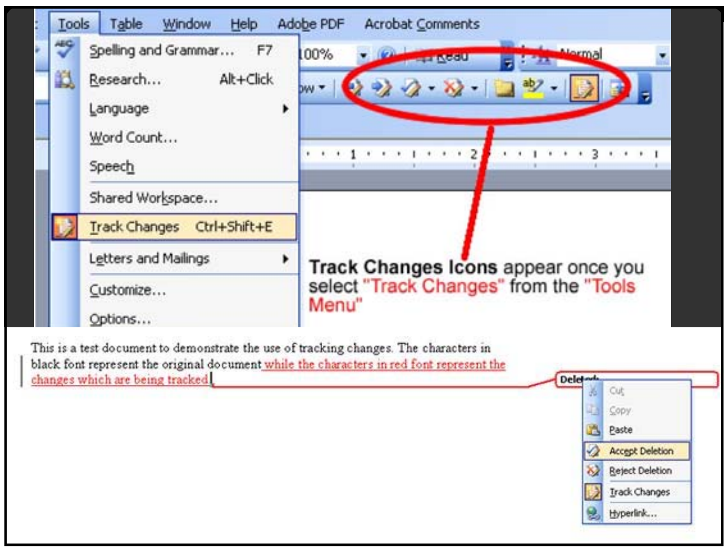
Final Project

Design a new visualization technique or system
Many options: new system, interaction technique, design study
6-8 page paper in conference paper format
2 Project Presentations

Schedule
Project Proposal: **Tuesday, Nov 15** *(end of day)*
Initial Presentation: **Tuesday, Nov 29**
Poster Presentation: **Tuesday, Dec 13** *(5-7pm)*
Final Papers: **Thursday, Dec 15** *(end of day)*

Logistics
Groups of up to 3 people, graded individually
Clearly report responsibilities of each member

Evolving Documents



Visualizing Revision History

How to depict contributions over time?

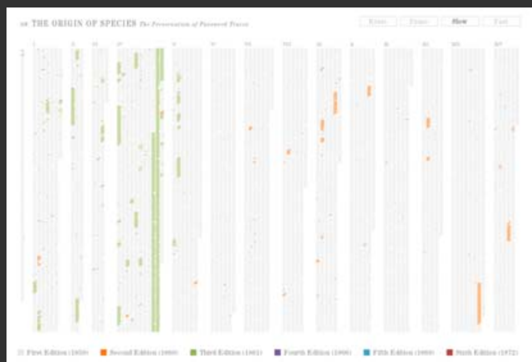
Example: Wikipedia history log

Chocolate
Revision history

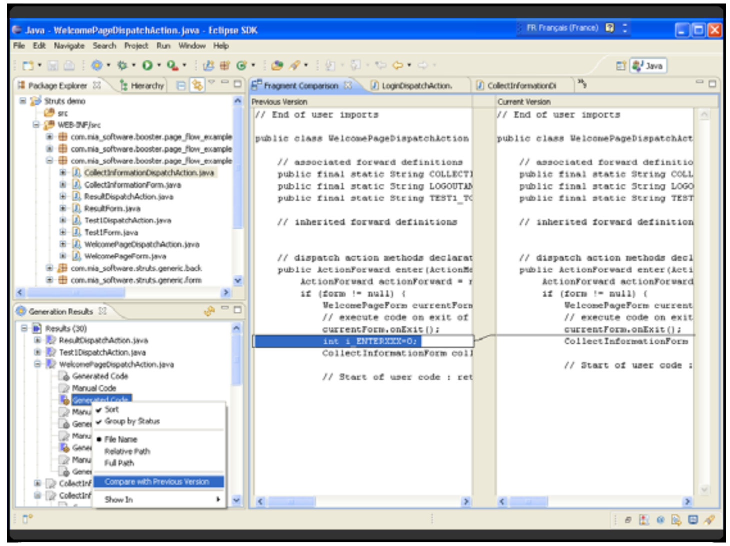
Legend: (cur) = difference with current version, (last) = difference with preceding version, M = minor edit

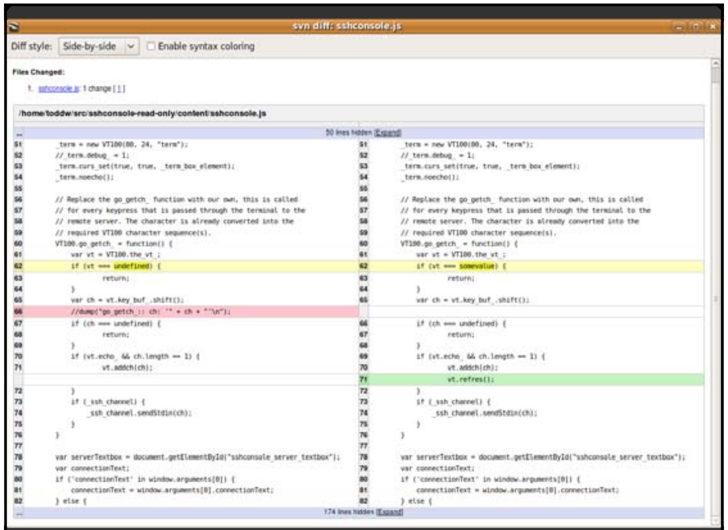
- (cur) (last) . . 12:01, 20 Aug 2003 . . Dysprosia *(neaten to do, rearrange see also)*
- (cur) (last) . . 11:59, 20 Aug 2003 . . Patrick
- (cur) (last) . . 11:52, 20 Aug 2003 . . 81.203.98.109
- (cur) (last) . . **M** 18:36, 6 Aug 2003 . . Manika *(corrected spelling)*
- (cur) (last) . . 18:32, 6 Aug 2003 . . Daniel Quinlan *(removing obscure heraldry information, belongs on [[heraldry]] if anywhere)*
- (cur) (last) . . 15:21, 6 Aug 2003 . . Rmhermen
- (cur) (last) . . 15:08, 6 Aug 2003 . . Cyp *(Chocolate often has odd shapes.)*
- (cur) (last) . . 19:14, 3 Aug 2003 . . Daniel C. Boyer *("chocolate" as shade of gules in heraldry)*
- (cur) (last) . . **M** 02:00, 30 Jul 2003 . . Evercat *(fmt)*

Animated Traces [Ben Fry]

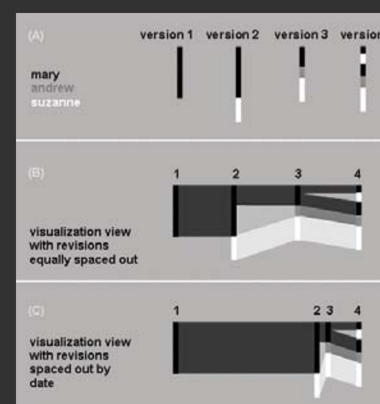


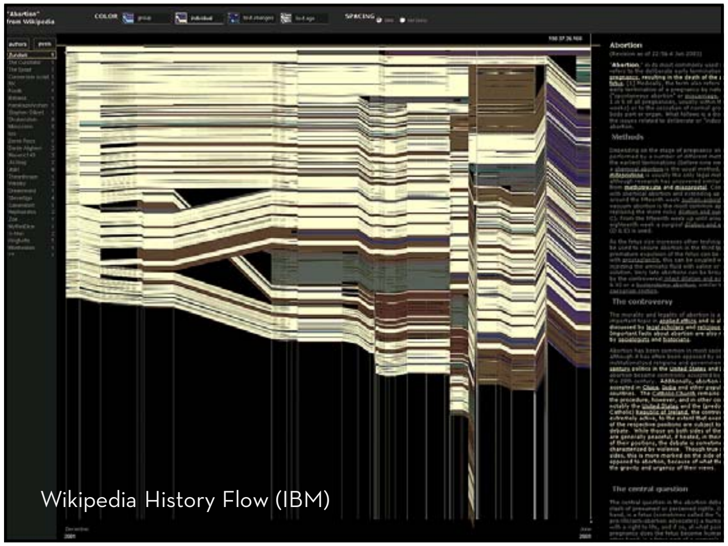
http://benfry.com/traces/



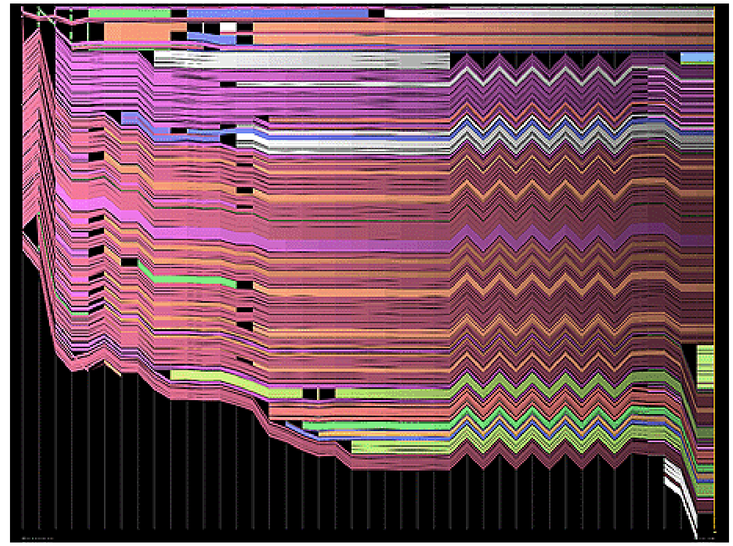


History Flow (Viégas et al)





Wikipedia History Flow (IBM)



Tips: Evolving documents

- High-level understanding
- Provide context
  - Show text within source document
  - Cross reference with other dimensions

Visualizing Conversation

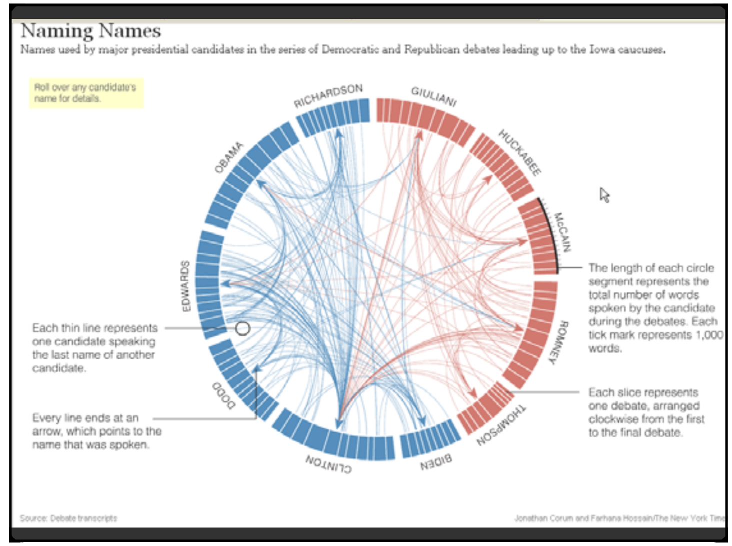
Visualizing Conversation

Many dimensions to consider:

- Who (senders, receivers)
- What (the content of communication)
- When (temporal patterns)

Interesting cross-products:

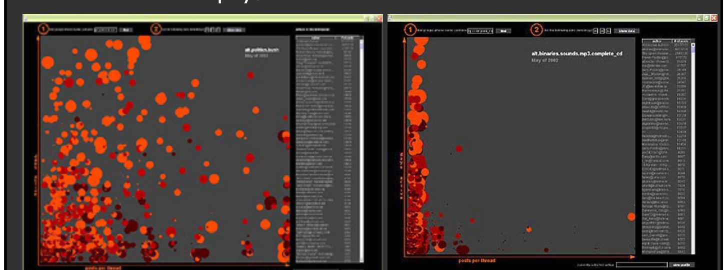
- What x When → Topic "Zeitgeist"
- Who x Who → Social network
- Who x Who x What x When → Information flow

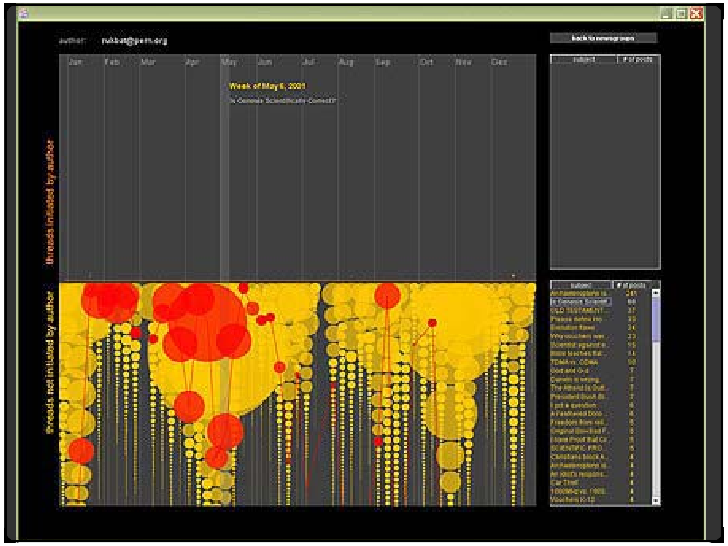


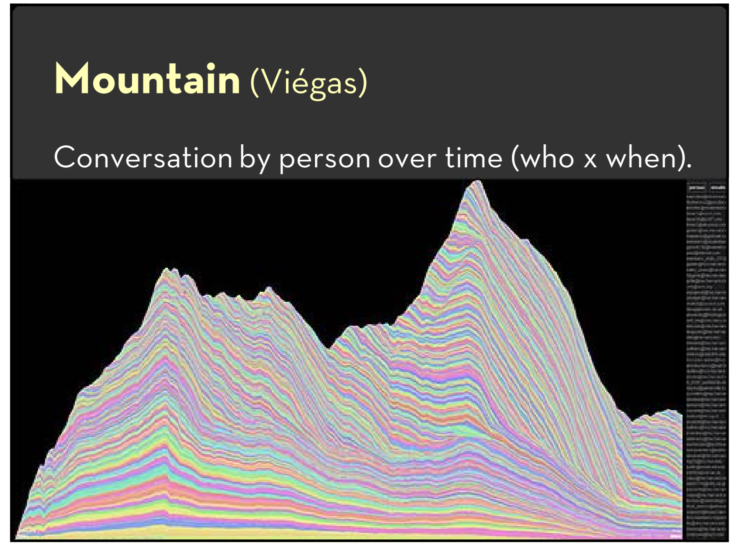
Usenet Visualization (Viégas & Smith)

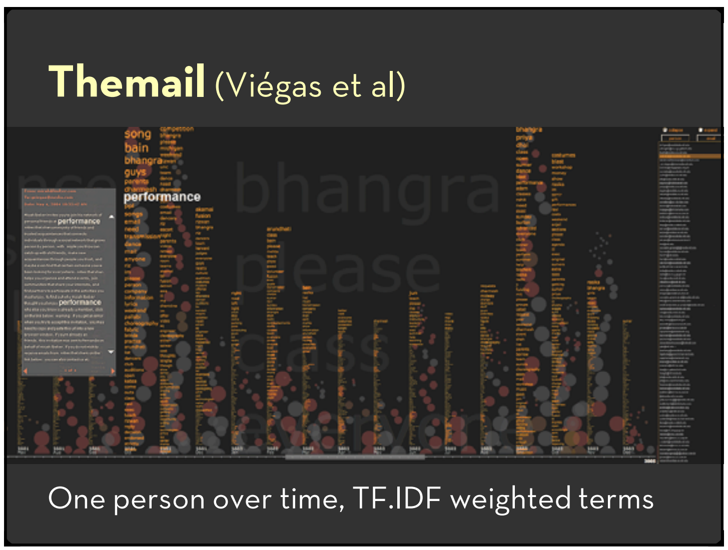
Show correspondence patterns in text forums

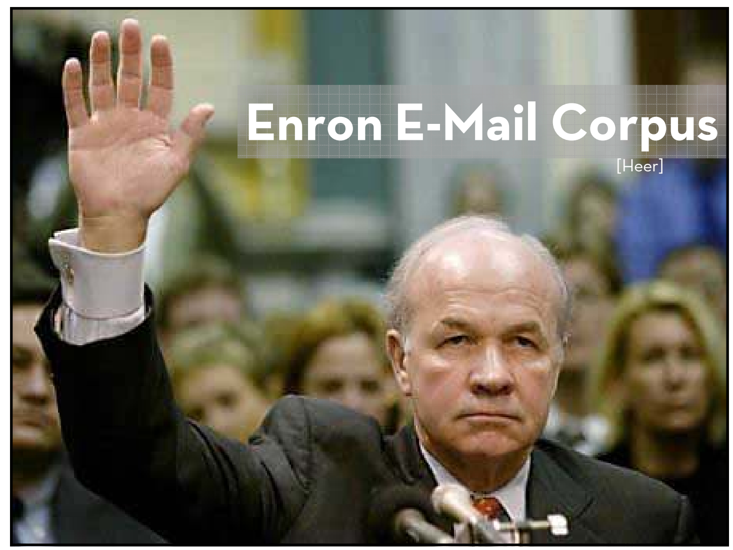
Initiate vs. reply; size and duration of discussion

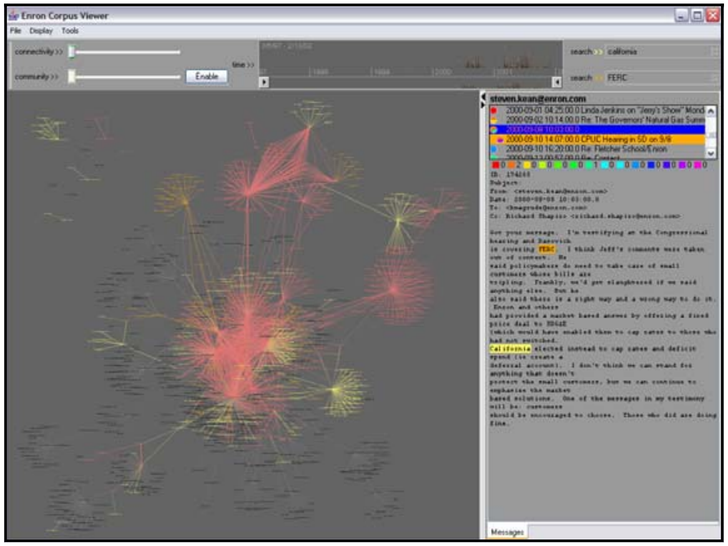


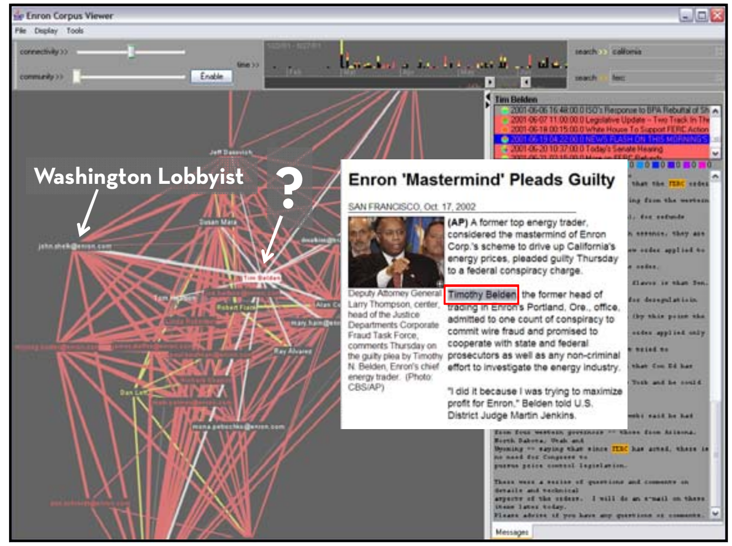



Mountain (Viégas)
Conversation by person over time (who x when).


Themail (Viégas et al)
One person over time, TF.IDF weighted terms


Enron E-Mail Corpus
[Heer]





Tips: Conversations

- Understand your units of analysis
  - Extract entities and relationships relevant to analysis task.
  - Cross-reference with other data dimensions.

Visualizing Document Collections

NewsMap: Google News Treemap (Marcos Weskamp)

10 x 10 News Map (Harris 2004)

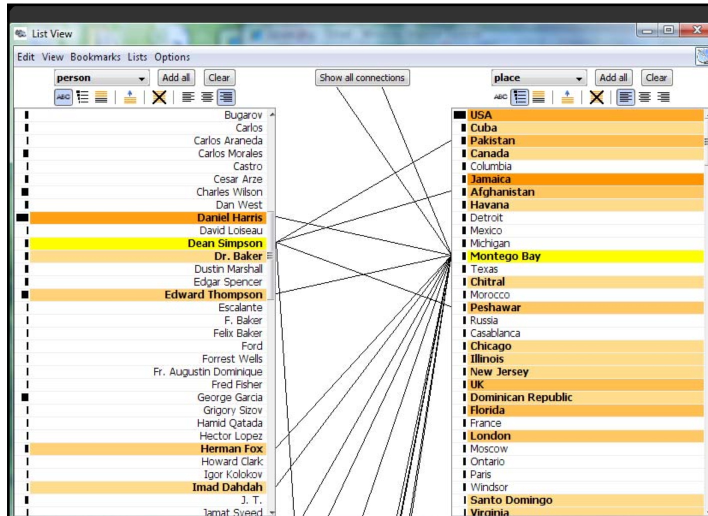
Named Entity Recognition

Identify and classify named entities in text:

- John Smith → PERSON
- Soviet Union → COUNTRY
- 353 Serra St → ADDRESS
- (555) 721-4312 → PHONE NUMBER

Entity relations: how do the entities relate?

- Simple approach: do they co-occur in a small window of text?



Doc. Similarity & Clustering

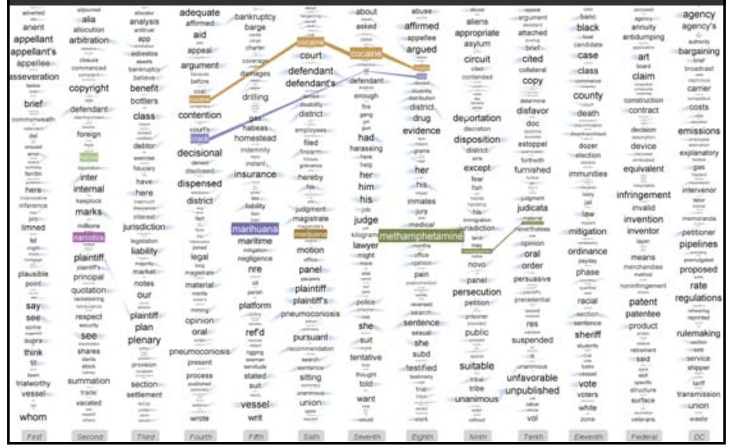
In vector model, compute distance among docs

- For TF.IDF, typically cosine distance
- Similarity measure can be used to cluster

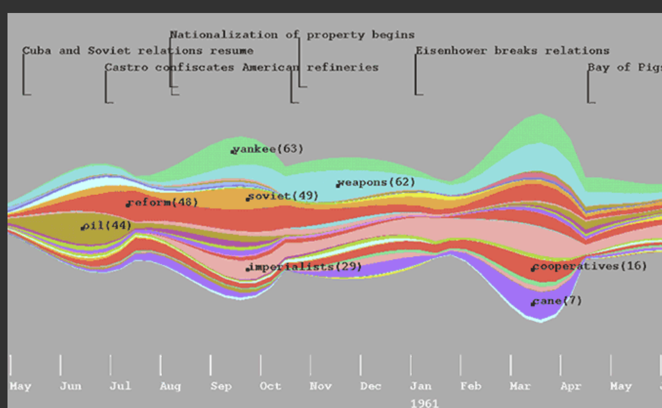
Topic modeling approaches

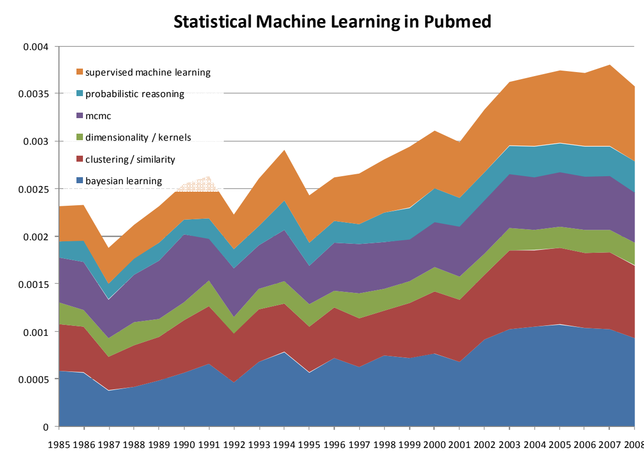
- Assume documents are a mixture of topics
- Topics are (roughly) a set of co-occurring terms
- Latent Semantic Analysis (LSA): reduce term matrix
- Latent Dirichlet Allocation (LDA): statistical model

Parallel Tag Clouds [Collins et al 09]



ThemeRiver [Havre et al 99]



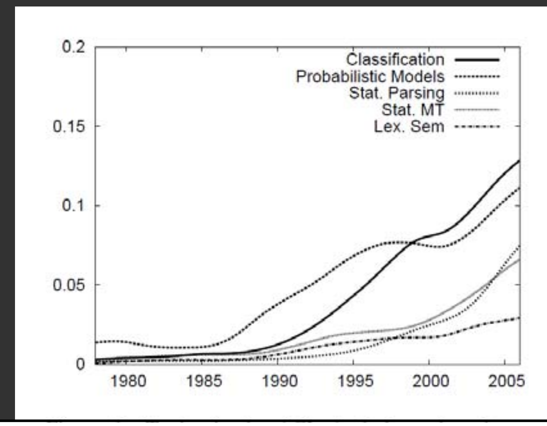


Track topic strengths over time

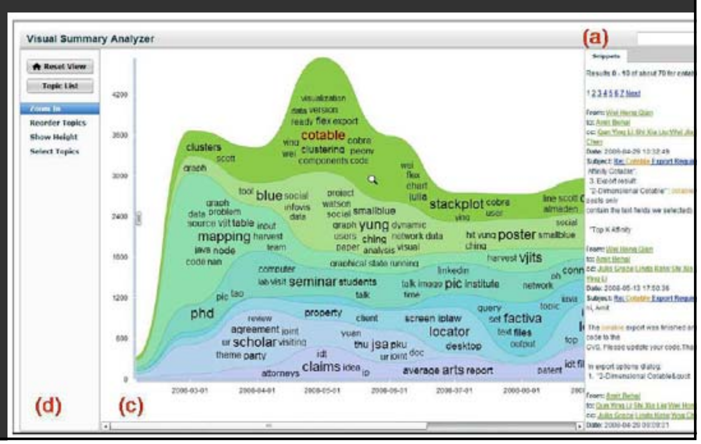
Interpretation and Trust?

- Interpretable topics?
- Trust the topics?

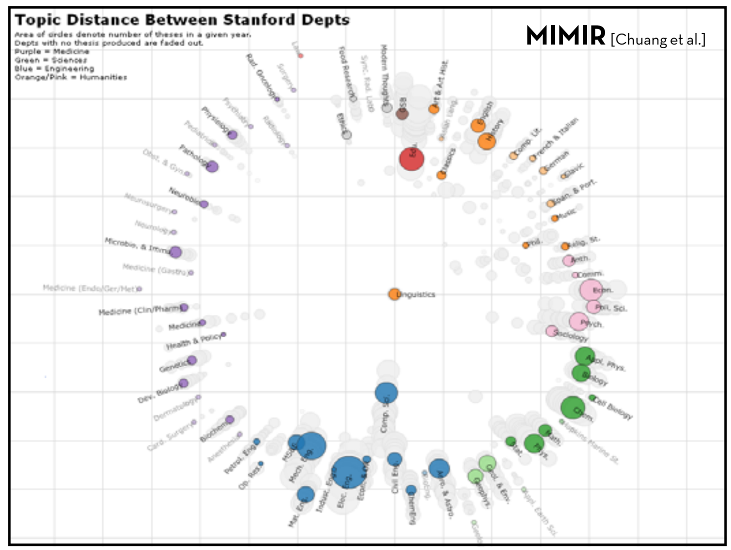
History of Comp Linguistics [Hall et al 06]



TIARA [Wei et al. 09]



MIMIR [Chuang et al.]



Challenges of Text Visualization

- High Dimensionality
  - Where possible use **text to represent text**…
    … which terms are the most descriptive?
- Context & Semantics
  - Provide **relevant context** to aid understanding.
  - Show (or provide access to) the **source text**.
- Modeling Abstraction
  - Determine your **analysis task**.
  - Understand abstraction of your **language models**.
  - Match analysis task with appropriate tools and models.

Lessons for Text Visualization

- Align analysis task with appropriate model.
- Handle high dimensionality…
  - Semantically
    - Interpretation: Longer phrases
    - Restaurant reviews: Adjective-noun word pairs
    - Relationships: Word sequences, hierarchy, clustering, …
    - Topic models: ***with care***
  - Visually
    - Word position within document
    - High-level structures in document collection
    - Visual representation matching semantic relationships

Lessons for Text Visualization

- Align analysis task with appropriate model.
- Provide context and semantics…
  - Apply appropriate text processing: stemming, named entities, etc.
  - Reverse stem for presentation
  - Show text within source document
  - Interaction to enable analysis cycle
  - Allow users to express contextual or domain knowledge
  - Cross-reference with other data dimensions