Text Visualization

Dr Puteri NE Nohuddin
IVI, UKM

Why Visualize Text?

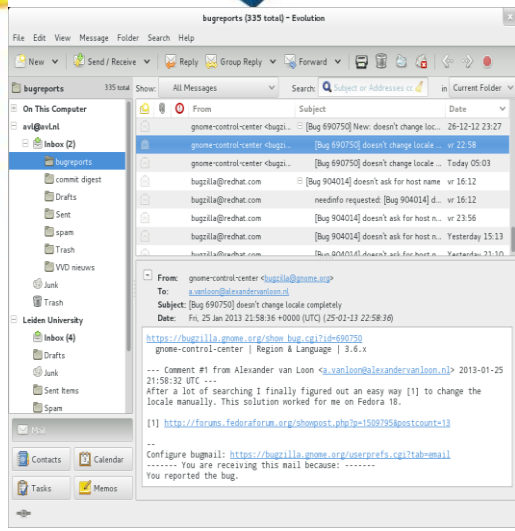
Understanding – get the "gist" of a document

Grouping – cluster for overview or classification

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

Correlation – 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

Collections of Documents

Messages (e-mail, blogs, tags, comments)

Social networks (personal profiles)

Academic collaborations (publications)

Example: Text Analysis and Visualization

Scenario of Malaysia's Economy

Text Data

News articles

Market reports

What questions might you want to answer?

What visualizations might help?

Tag Clouds: Word Count

Malaysia Economy

Word Tree: Word Sequences

Unstructured Data and Degree of Evaluation

Many text visualizations do not represent the text directly. They represent the output of a **language model** (word counts, word sequences, term links, term trends, 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?

Text Visualization Challenges

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.

Text as Data

Words as nominal data?

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 Pre-processing

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.*

Text Pre-processing

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*

Text Pre-processing

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

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

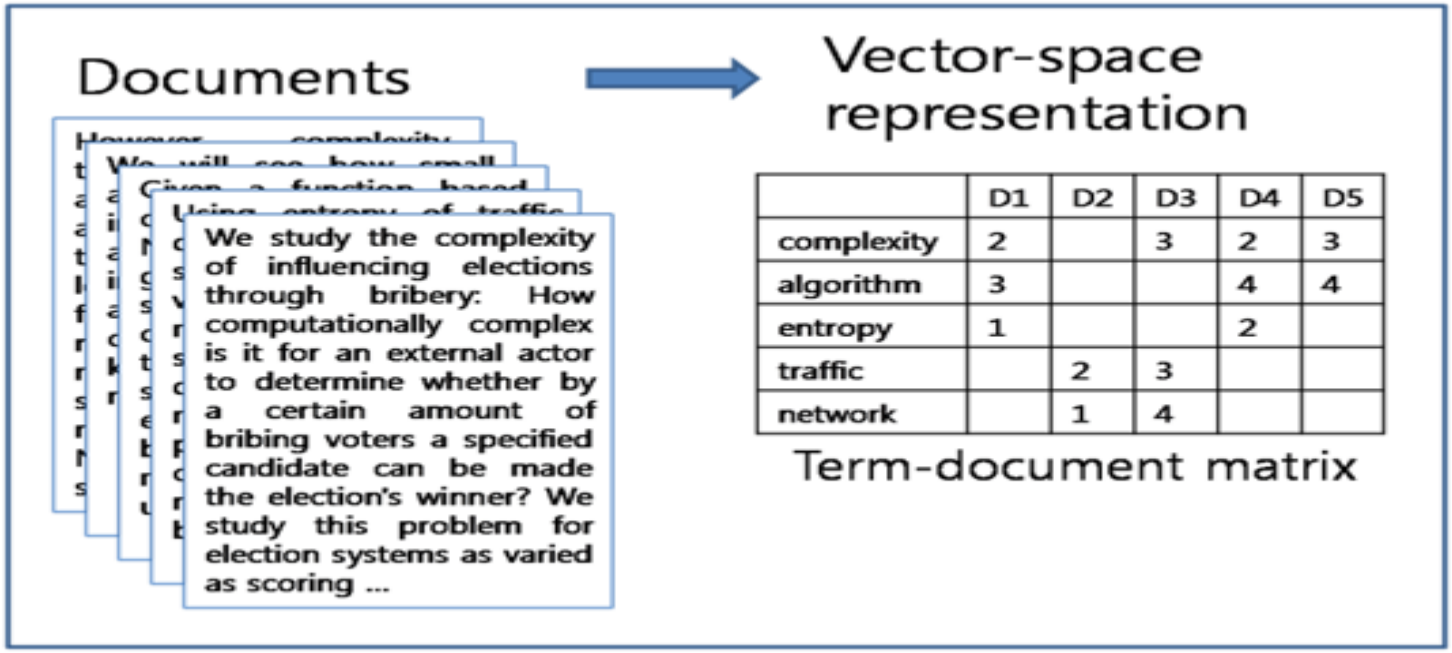
Aggregate into a document-term matrix

- Document vector space model

Document-Term Matrix

Each document is a vector of term weights

Simplest weighting is to just count occurrences



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Visualizations : Wordle of Sarah Palin RNC 9/3/2008 Speech

Creator: Anonymous
Tags:

Edit Language Font Layout Color

Data file: Sarah Palin speaks at the Republican National Convention, 9/3/2008 Data source: SFGate / AP This data set has not yet been rated

Tag Clouds or Word Clouds

Strengths

Can help with gisting and 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

Tips: Descriptive Phrases

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

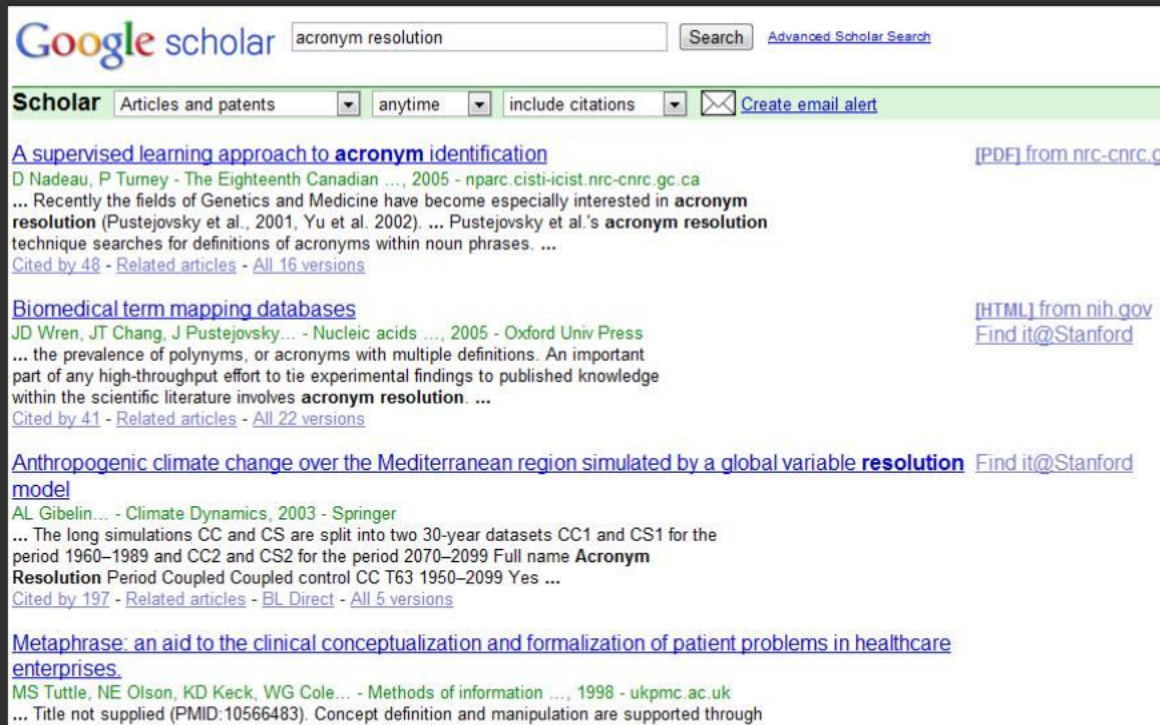
Document Content

Information Retrieval

Search for documents

Match query string with documents

Visualization to **contextualize results**



THE WORDS THAT WERE USED

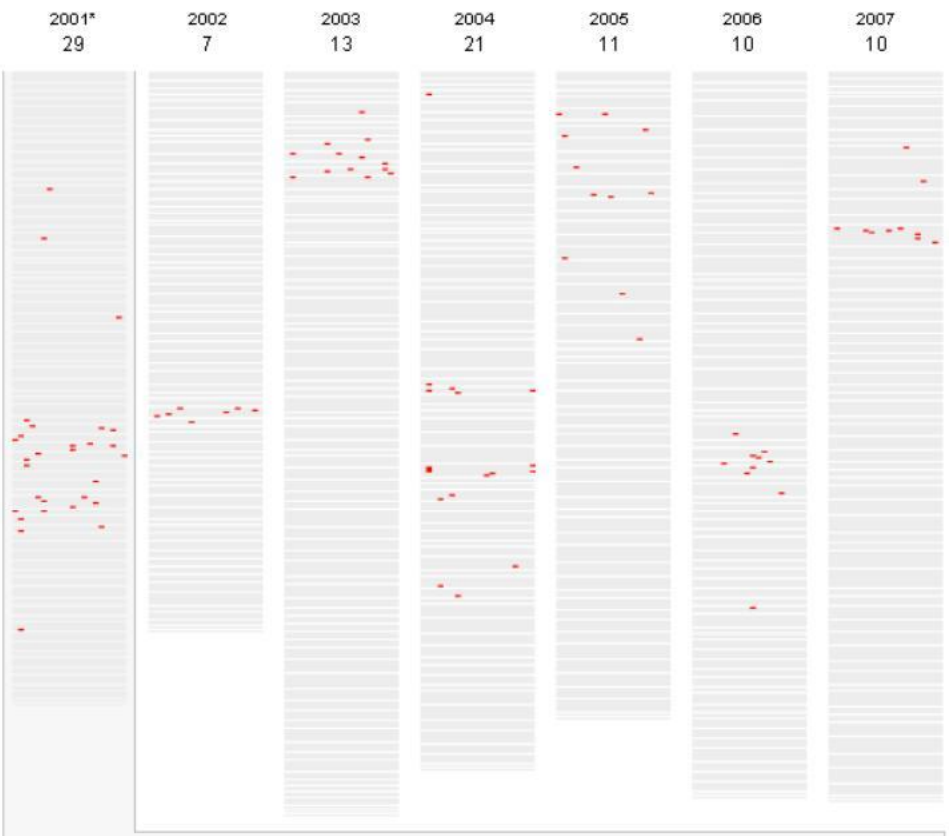
READ 2007 SPEECH | FEEDBACK

The 2007 State of the Union Address

Over the years, President Bush's State of the Union address has averaged almost 5,000 words each, meaning the the President has delivered over 34,000 words. Some words appear frequently while others appear only sporadically. Use the tools below to analyze what Mr. Bush has said.

Search | or choose a word here.

Use of the phrase "Tax" in past State of the Union Addresses



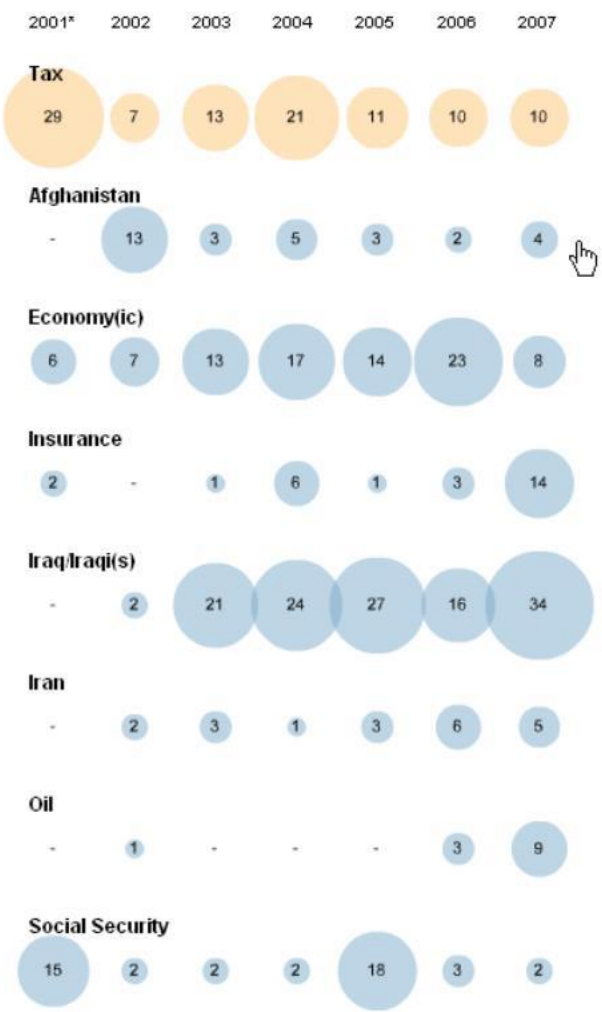
Next Instance of 'Tax'

The word in context

I believe in local control of schools. We should not, and we will not, run public schools from Washington, D.C. Yet when the federal government spends **TAX** dollars, we must insist on results. Children should be tested on basic reading and math skills every year between grades three and eight. Measuring is the only way to know whether all our children are learning. And I want to know, because I refuse to leave any child behind in America.

-- 2001 (Paragraph 14 of 73)

Compared with other words



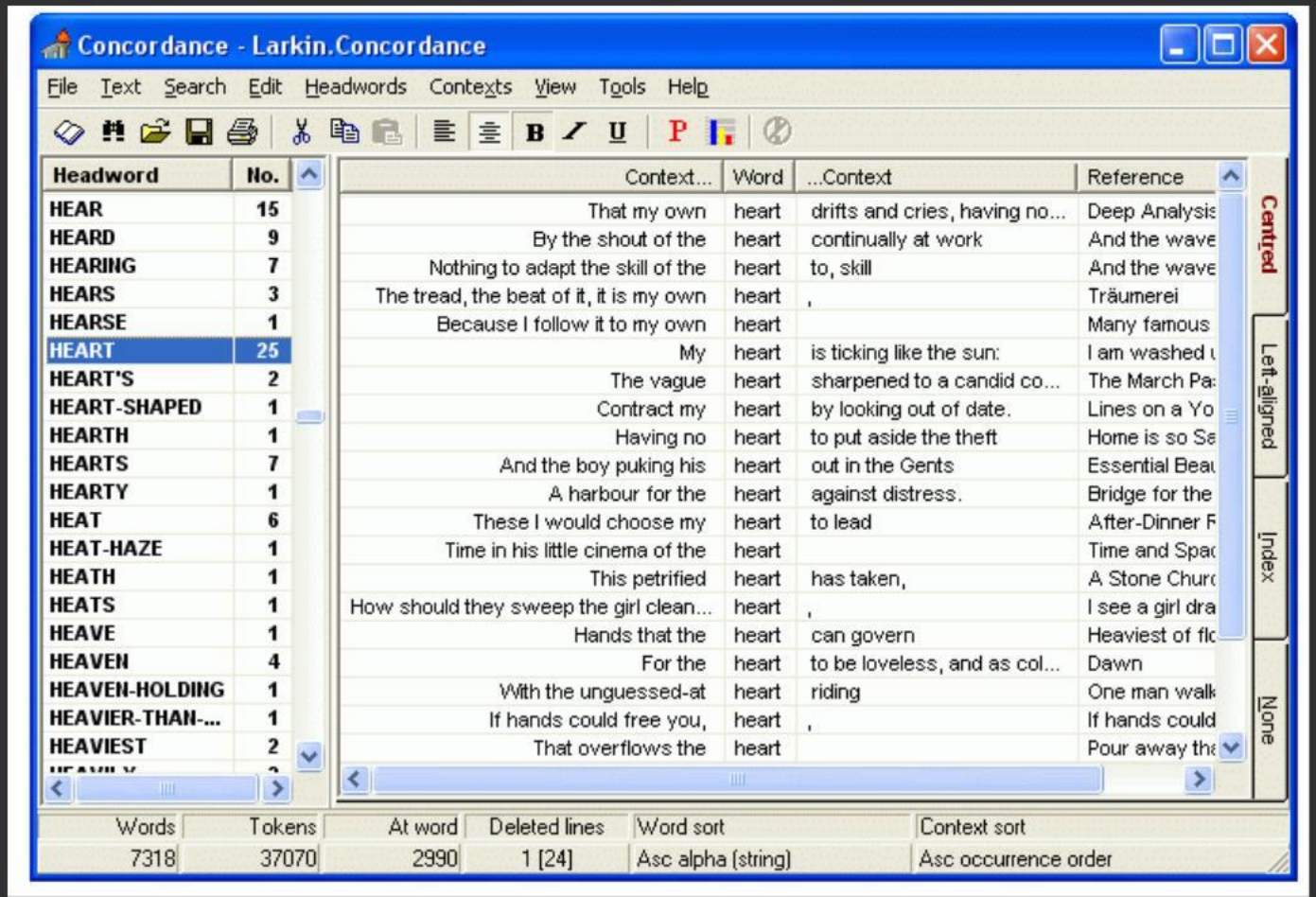
* As a newly elected president, Mr. Bush did not deliver a formal State of the Union address in 2001. His Feb. 27 speech to a joint session of Congress was analogous to the State of the Union, but without the title.

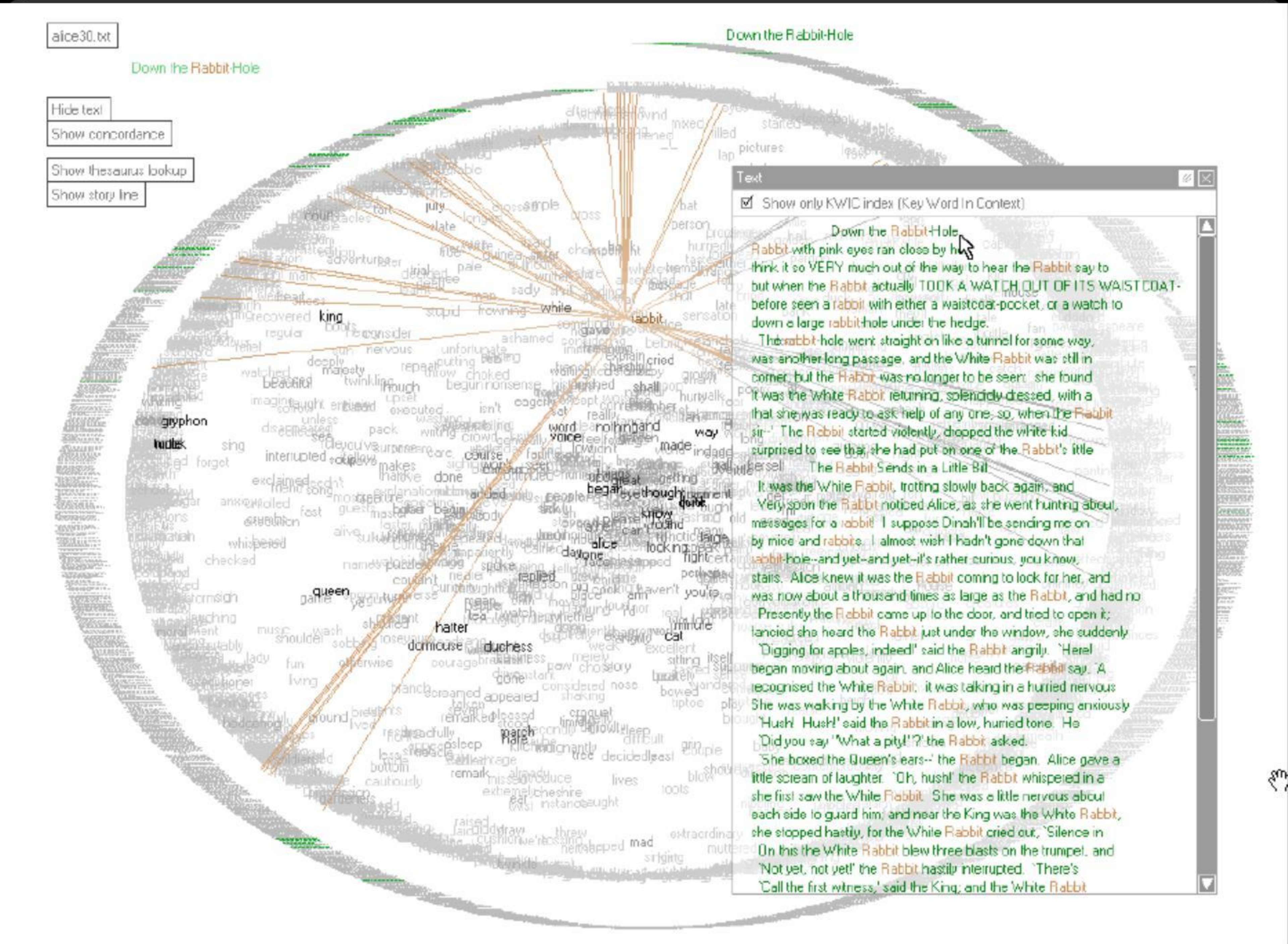
Ben Werschkul/The New York Times

New York Times

Concordance

What is the common local context of a term?

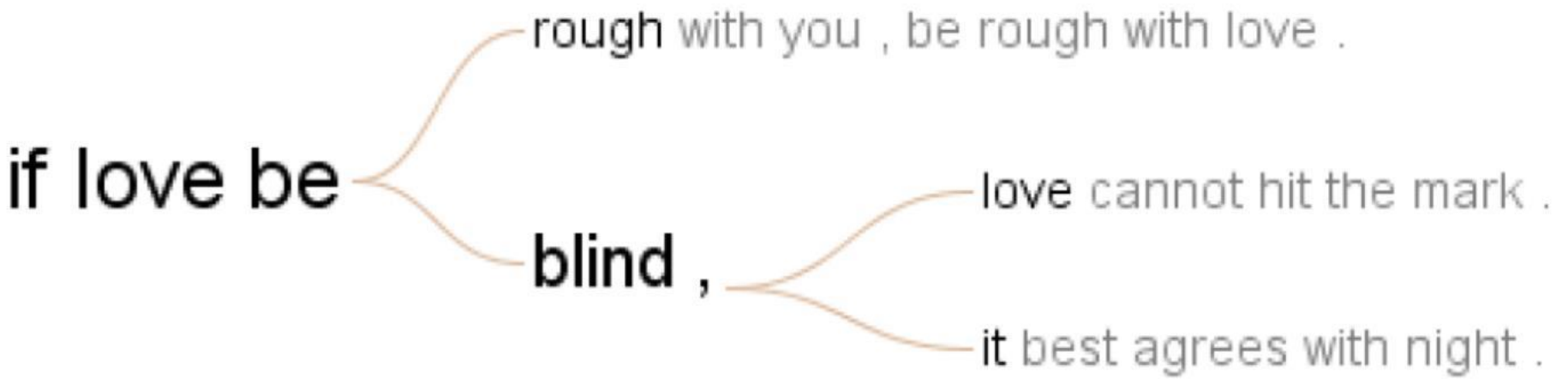



alice30.txt
Down the Rabbit-Hole
Hide text
Show concordance
Show thesaurus lookup
Show story line
Down the Rabbit-Hole
Text
Show only KWIC index (Key Word In Context)
Down the Rabbit-Hole
Rabbit with pink eyes ran close by her
think it so VERY much out of the way to hear the Rabbit say to
but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-
before seen a rabbit with either a waistcoat-pocket, or a watch to
down a large rabbit-hole under the hedge.
The rabbit-hole went straight on like a tunnel for some way,
was another long passage, and the White Rabbit was still in
corner, but the Rabbit was no longer to be seen: she found
It was the White Rabbit returning, splendidly dressed, with a
that she was ready to ask help of any one; so, when the Rabbit
sir--' The Rabbit started violently, dropped the white kid
surprised to see that she had put on one of the Rabbit's little
The Rabbit Sends in a Little Bill
It was the White Rabbit, trotting slowly back again, and
Very soon the Rabbit noticed Alice, as she went hunting about,
messages for a rabbit! I suppose Dinah'll be sending me on
by mice and rabbits. I almost wish I hadn't gone down that
rabbit-hole--and yet--and yet--it's rather curious, you know,
stairs. Alice knew it was the Rabbit coming to look for her, and
was now about a thousand times as large as the Rabbit, and had no
Presently the Rabbit came up to the door, and tried to open it;
fancied she heard the Rabbit just under the window, she suddenly
`Digging for apples, indeed!' said the Rabbit angrily. `Here!
began moving about again, and Alice heard the Rabbit say, `A
recognised the White Rabbit: it was talking in a hurried nervous
She was walking by the White Rabbit, who was peeping anxiously
`Hush! Hush!' said the Rabbit in a low, hurried tone. He
`Did you say "What a pity!"?' the Rabbit asked.
`She boxed the Queen's ears--' the Rabbit began. Alice gave a
little scream of laughter. `Oh, hush!' the Rabbit whispered in a
she first saw the White Rabbit. She was a little nervous about
each side to guard him; and near the King was the White Rabbit,
she stopped hastily, for the White Rabbit cried out, `Silence in
On this the White Rabbit blew three blasts on the trumpet, and
`Not yet, not yet!' the Rabbit hastily interrupted. `There's
`Call the first witness,' said the King; and the White Rabbit

if love be rough with you , be rough with love .

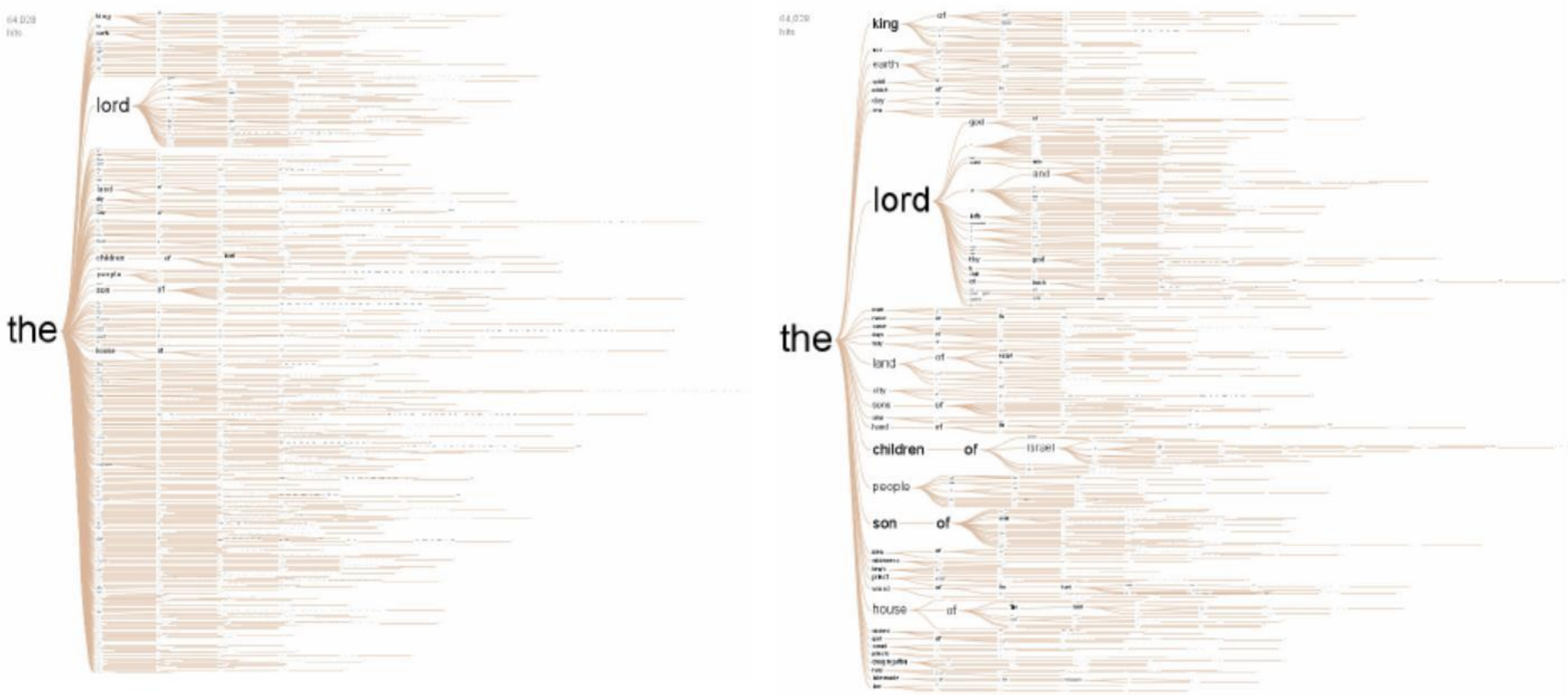
if love be blind , love cannot hit the mark .

if love be blind , it best agrees with night .

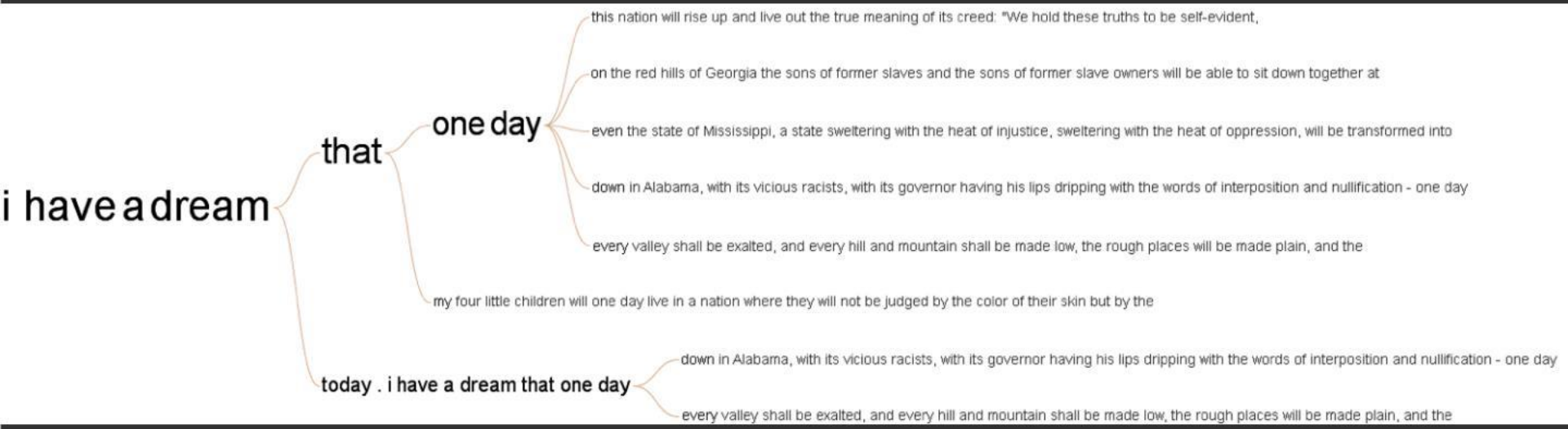


Word Tree [Wattenberg et al.]

Filter Infrequent Runs



Recurrent Themes in Speeches



Visualizations : Word tree / Alberto Gonzales

Creator: Martin Wattenberg
Tags:

explore
visualizations
data sets
comments
topic hubs

participate
create visualization
upload data set
create topic hub
register

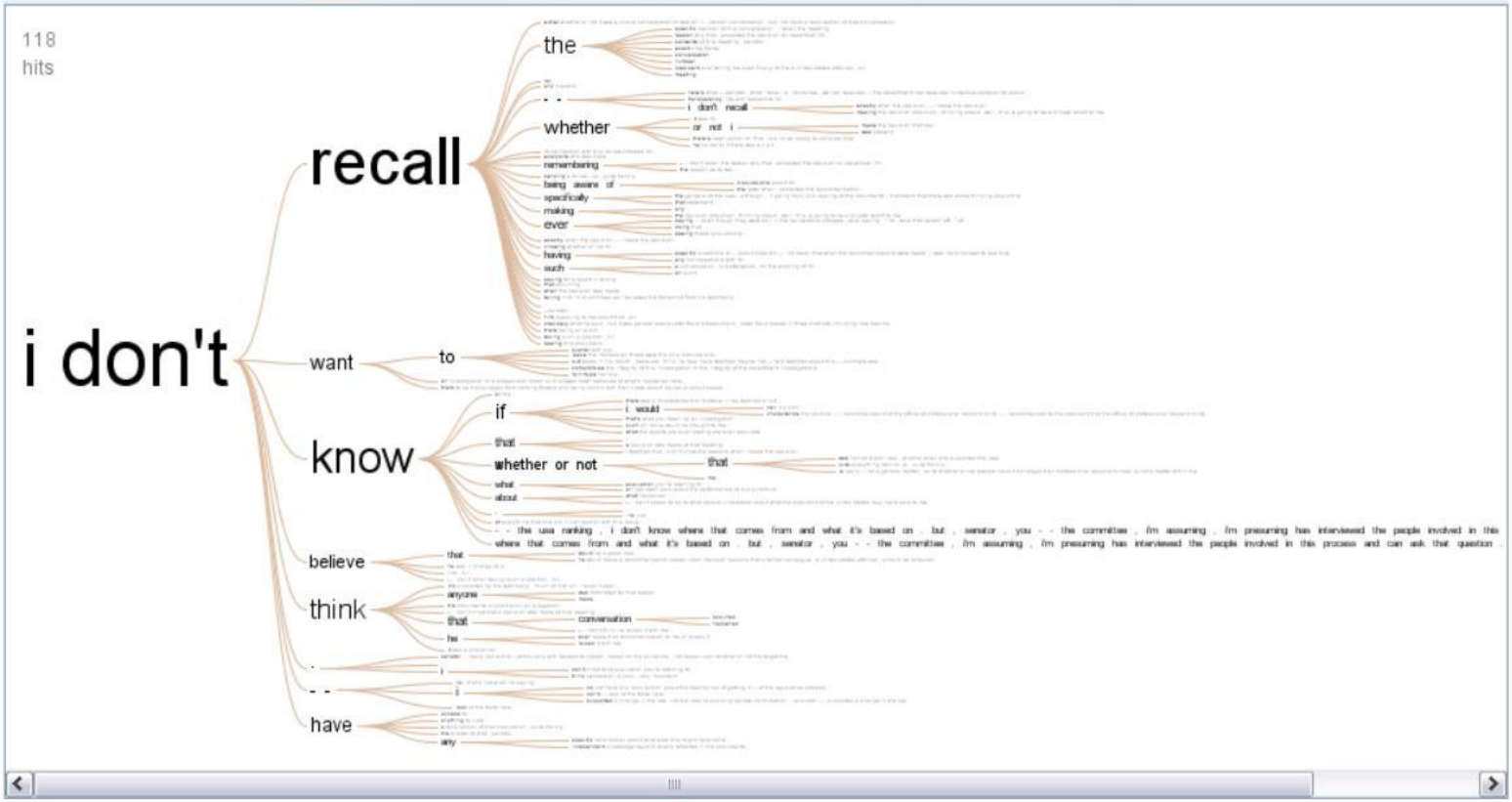
learn more
quick start
visualization types
data format & style
about Many Eyes
FAQ
blog

contact Us
contact
report a bug

legal
terms of use

Popular Dataset Tags
2007 2008 bible blog books census crime education eharmony election energy food health inauguration internet ireland literature lyrics media music network obama people politics population president prices religion

Search i don't | Back | Forward | Start | End | Occurrence Order | Clicks Will Zoom



Data file: Word in testimony from Gonzales, 4/19/2007 Data source: CQ Transcript Wire via the Washington Post This data set has not yet been rated

full image | share this | watch this | add to topic hub | rate this

Comments (4)

currently showing

This visualization has 4 positive and 0 negative

Glimpses of Structure…

Concordances show local, repeated structure

But what about other types of patterns?

Lexical: <A> at <B>

Syntactic: <Noun> <Verb> <Object>

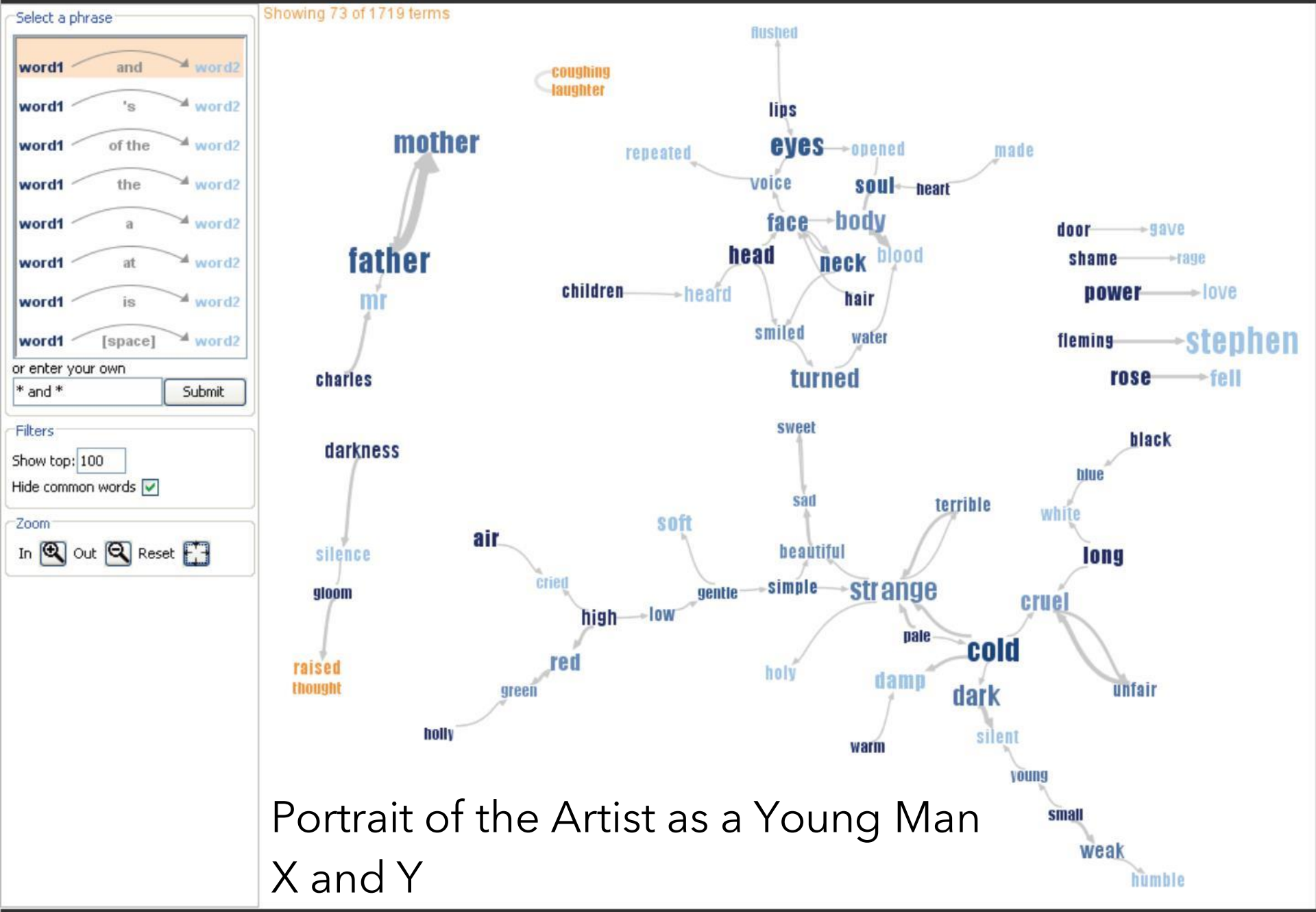
Phrase Nets [van Ham et al.]

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 patterns in a node-link view

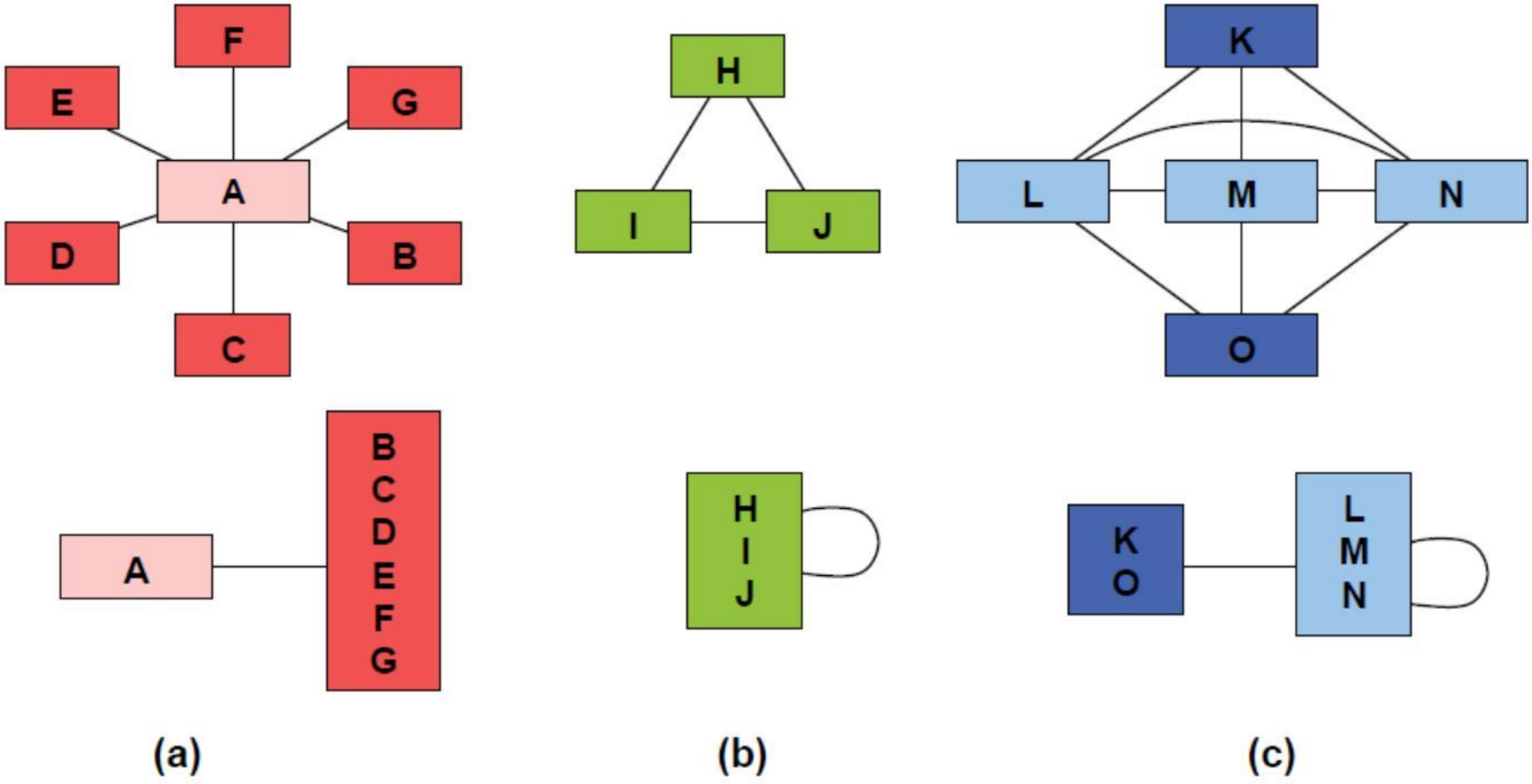
- Occurrences -> Node size
- Pattern position -> Edge direction



Portrait of the Artist as a Young Man

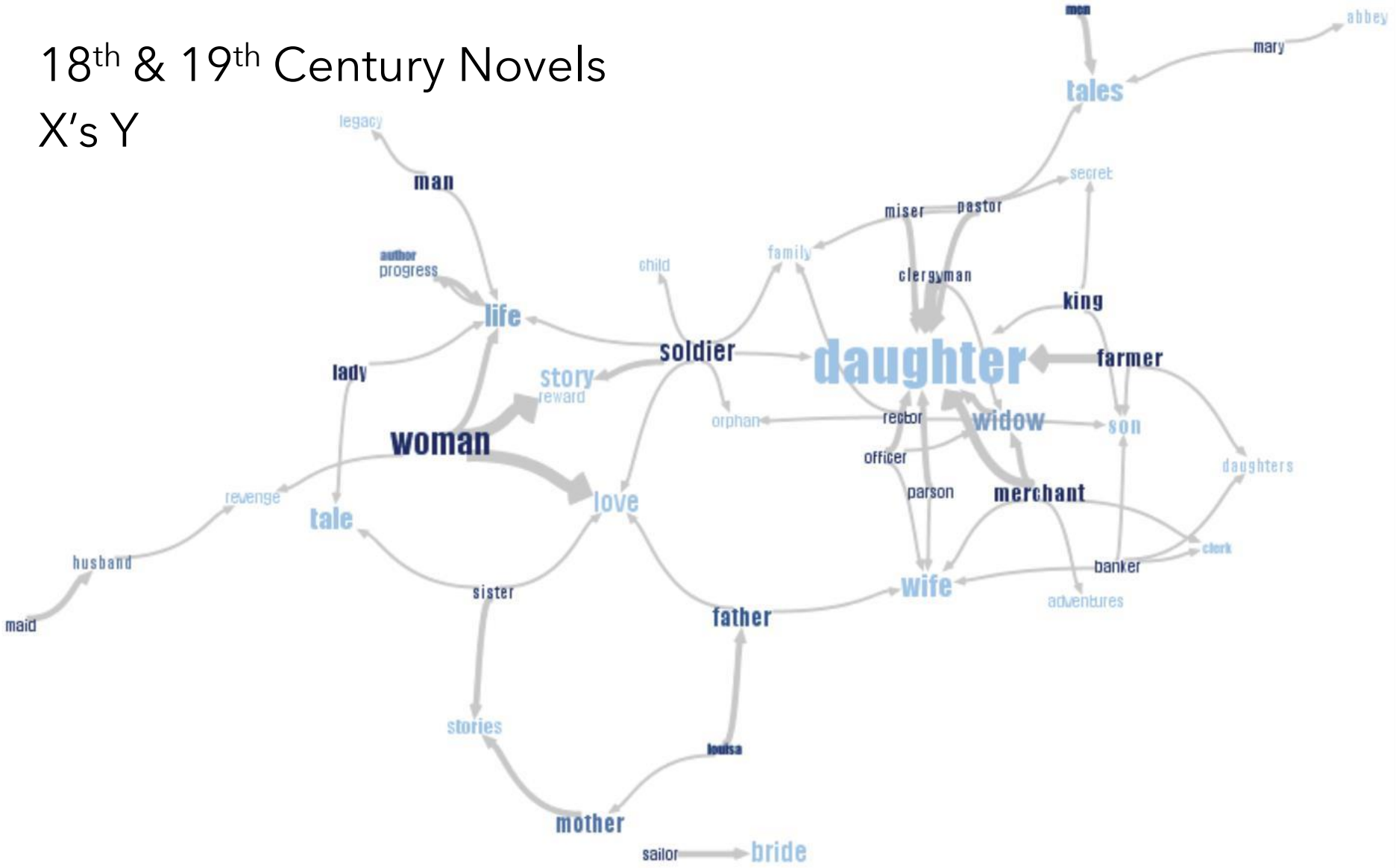
X and Y

Node Grouping



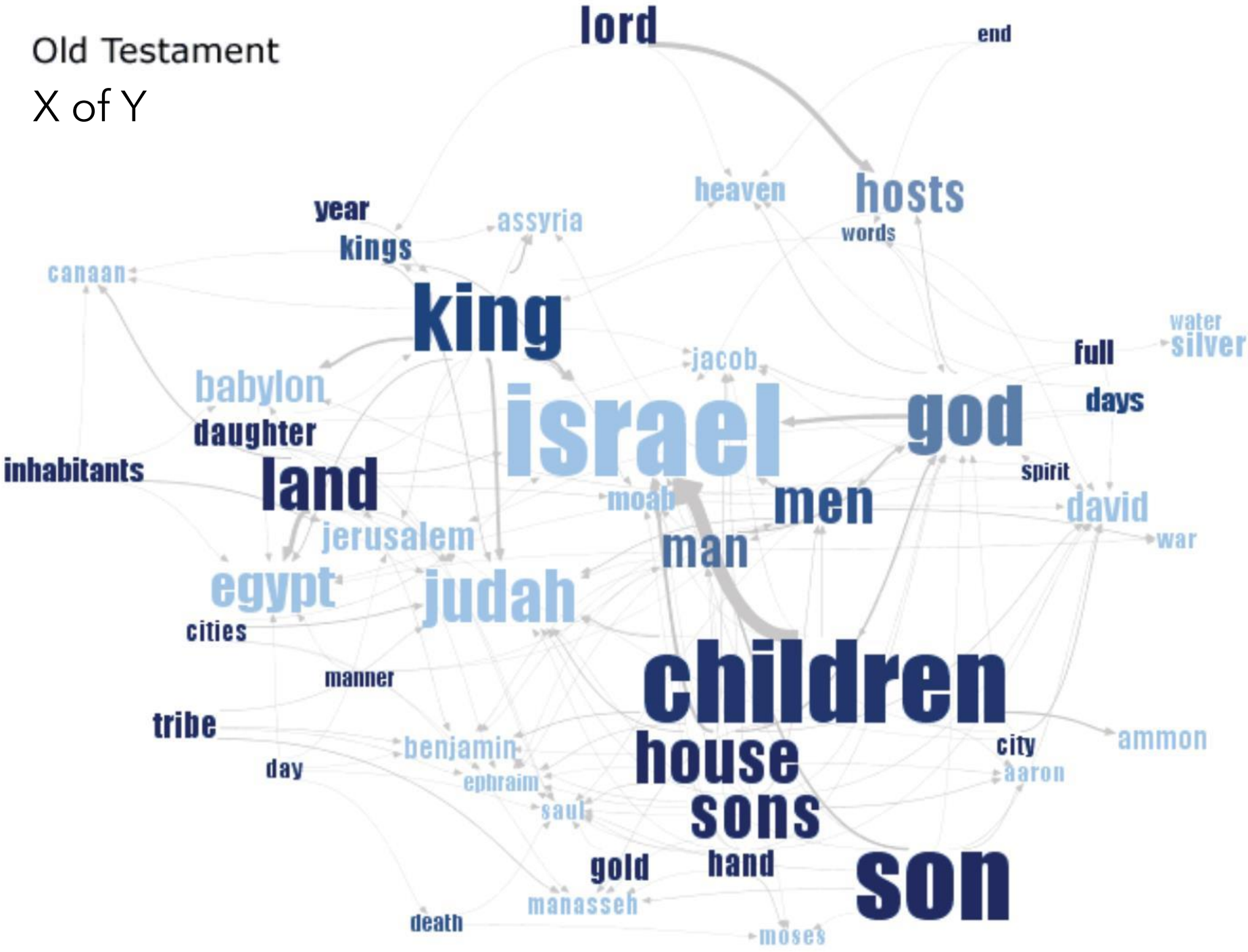
18th & 19th Century Novels

X's Y



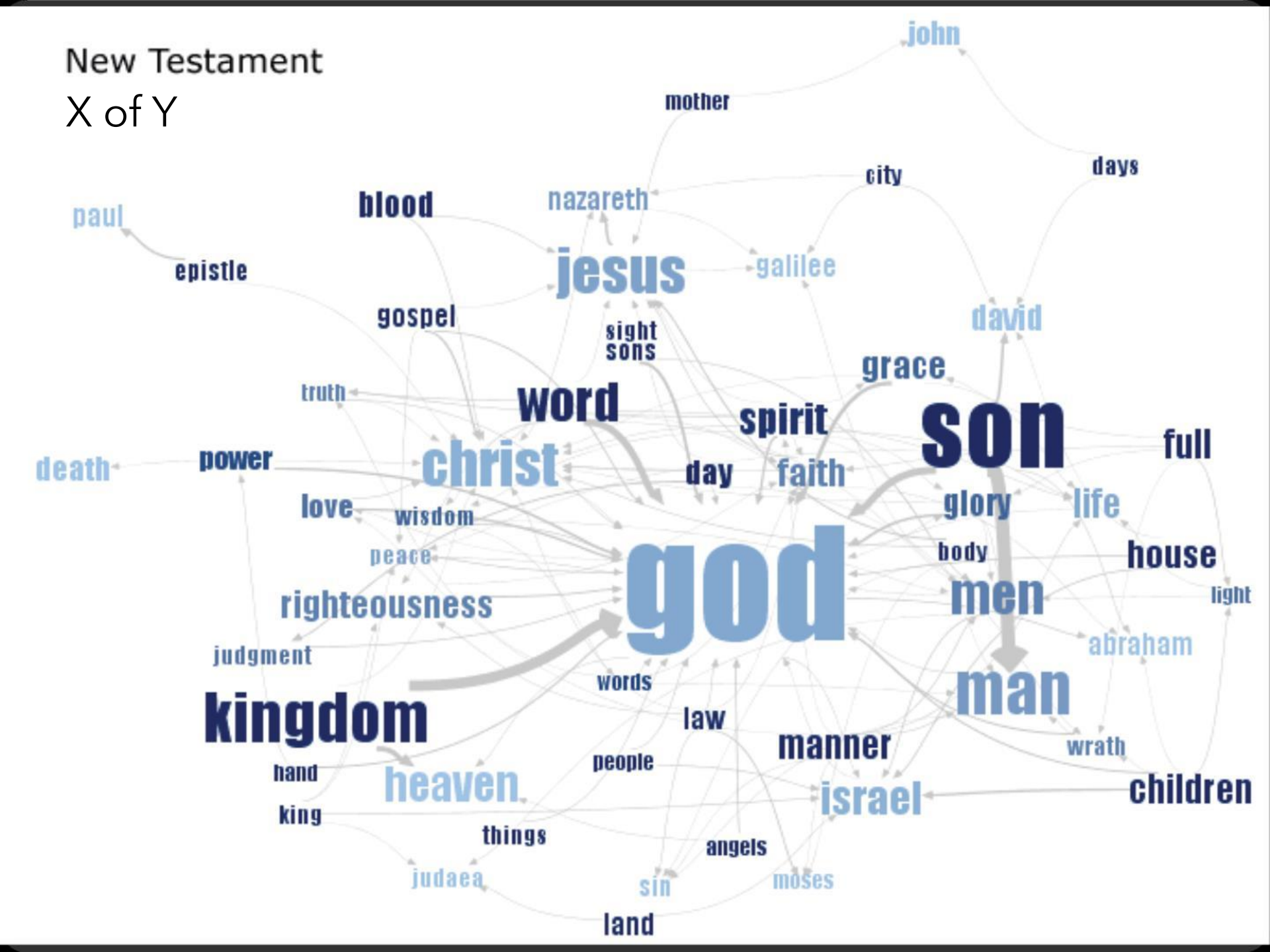
Old Testament

X of Y



New Testament

X of Y



Document Content

Understand Your Analysis Task

Visually: Word position, browsing, brush & link

Semantically: Word sequence, hierarchy, clustering

Both: Spatial layout reflects semantic relationships

The Role of Interaction

Language model supports visual analysis cycles

Allow modifications to the model: custom patterns for expressing contextual or domain knowledge

Conversations

Visualizing Conversation

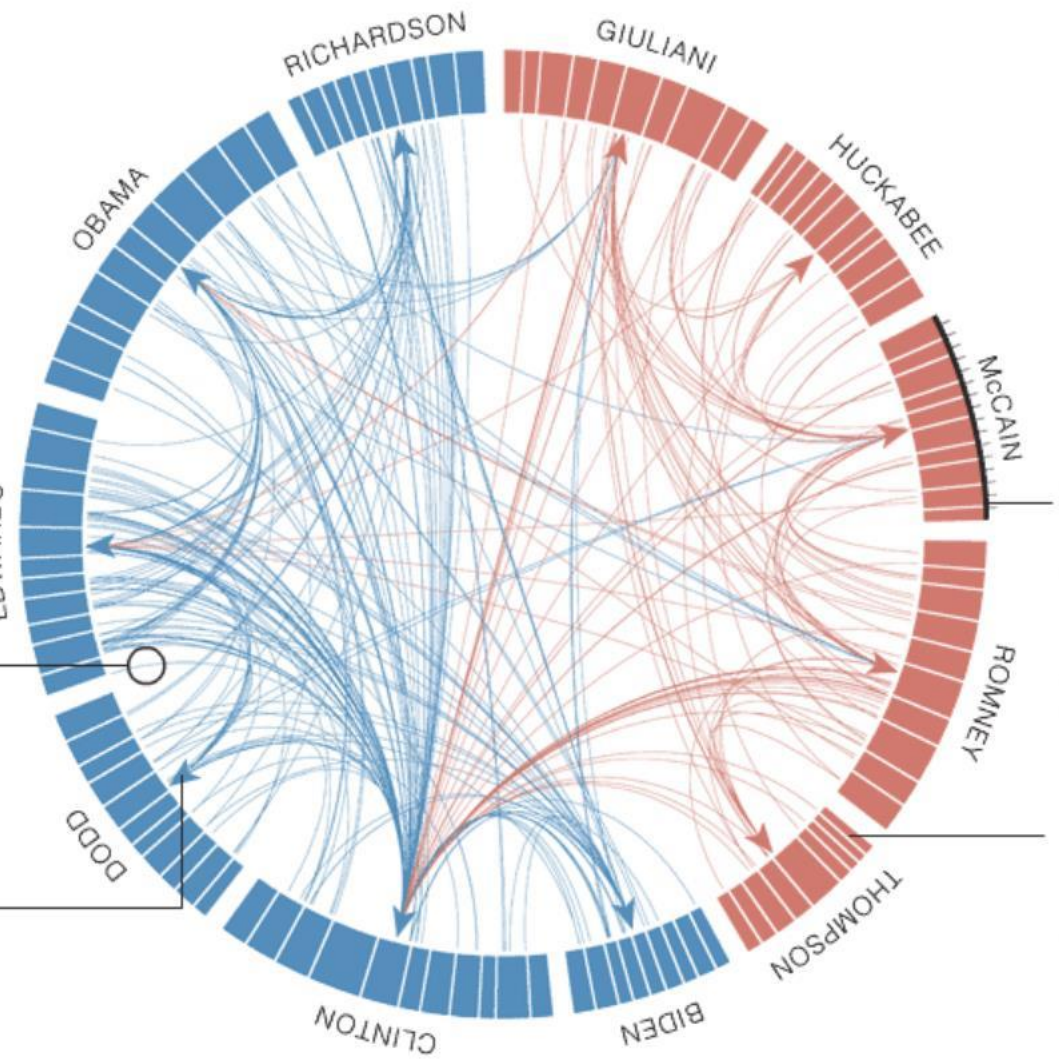
Many dimensions to consider:
Who (senders, receivers)
What (the content of communication)
When (temporal patterns)

Interesting cross-products:
What x When -> Topic “Zeitgeist”
Who x Who -> Social network
Who x Who x What x When -> Information flow

Naming Names

Names used by major presidential candidates in the series of Democratic and Republican debates leading up to the Iowa caucuses.

Roll over any candidate's name for details.



Each thin line represents one candidate speaking the last name of another candidate.

Every line ends at an arrow, which points to the name that was spoken.

The length of each circle segment represents the total number of words spoken by the candidate during the debates. Each tick mark represents 1,000 words.

Each slice represents one debate, arranged clockwise from the first to the final debate.

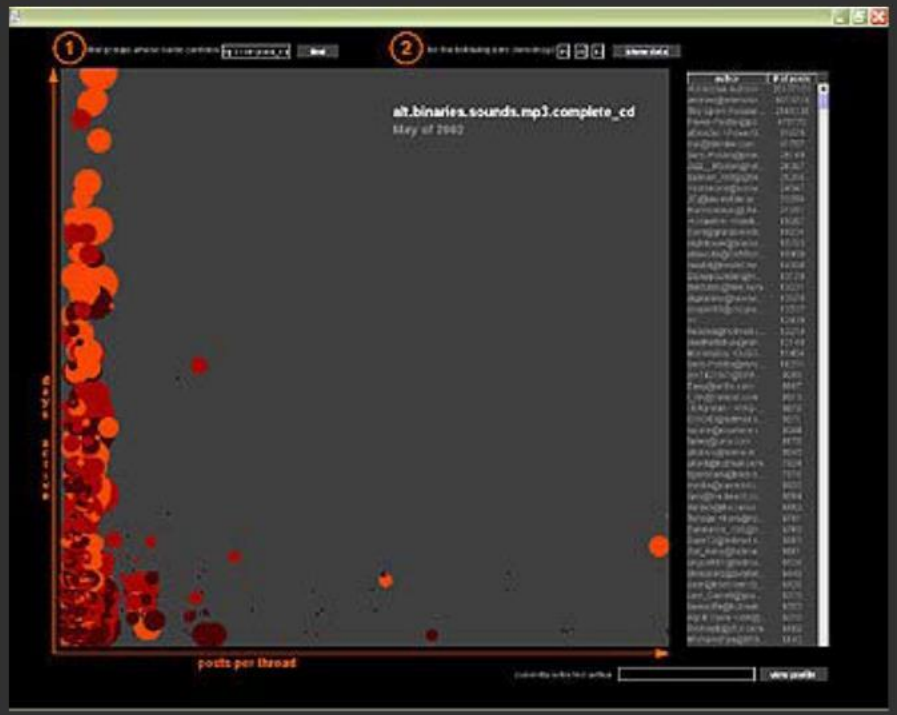
Source: Debate transcripts

Jonathan Corum and Farhana Hossain/The New York Times

Usenet Visualization [Viegas & Smith]

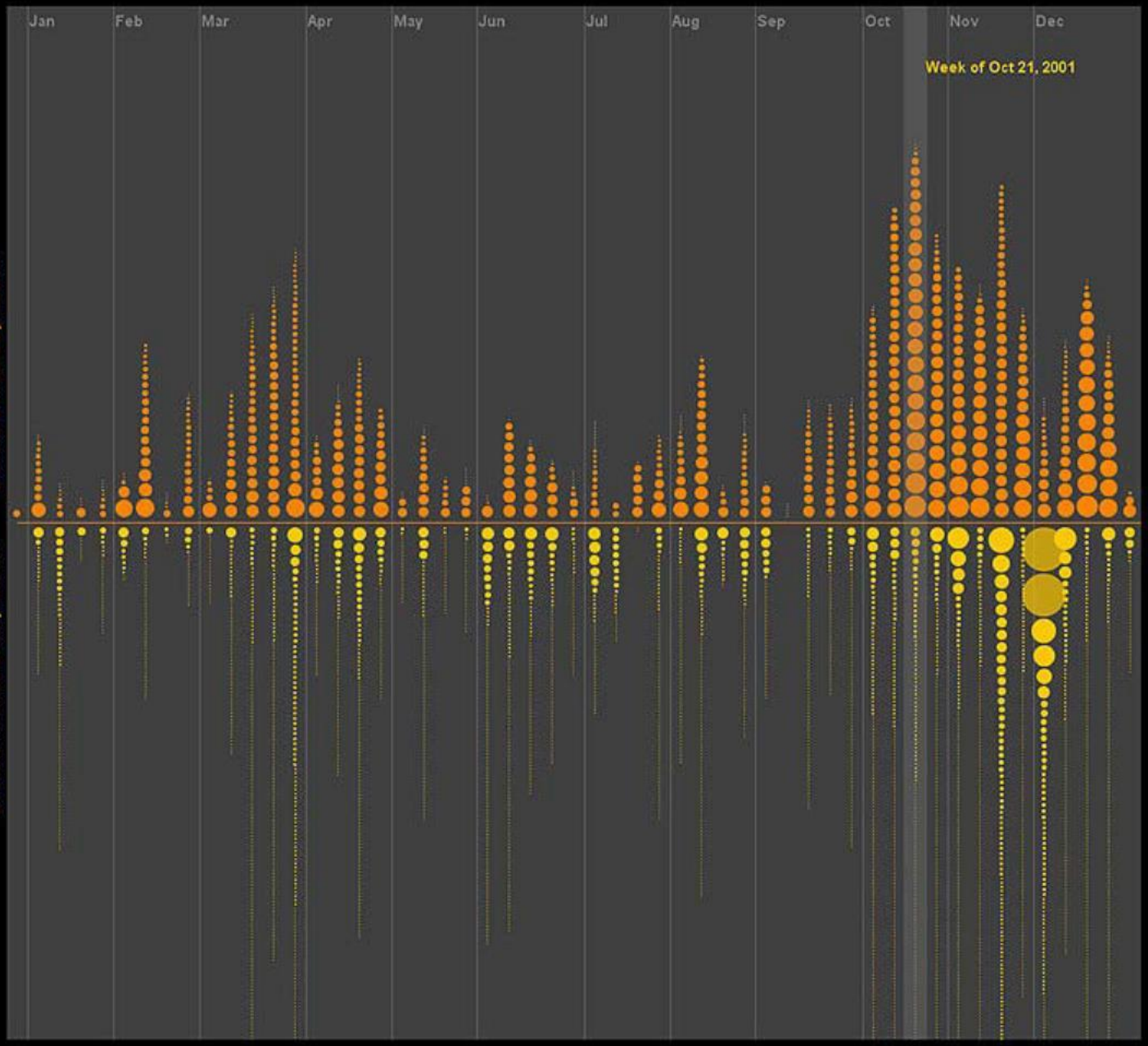
Show correspondence patterns in text forums

Initiate vs. reply; size and duration of discussion



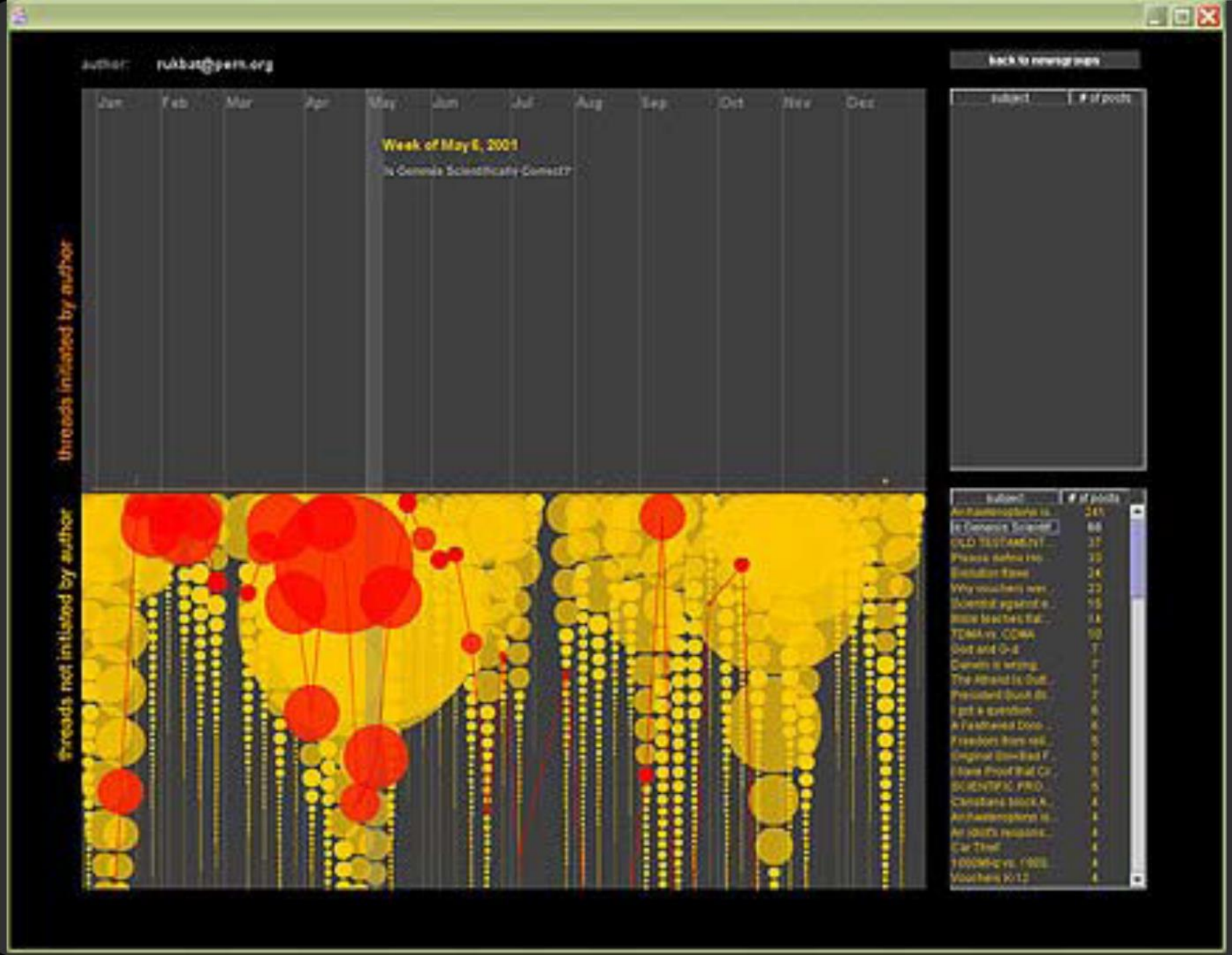
author: jillyb@mail.com

back to newsgroups



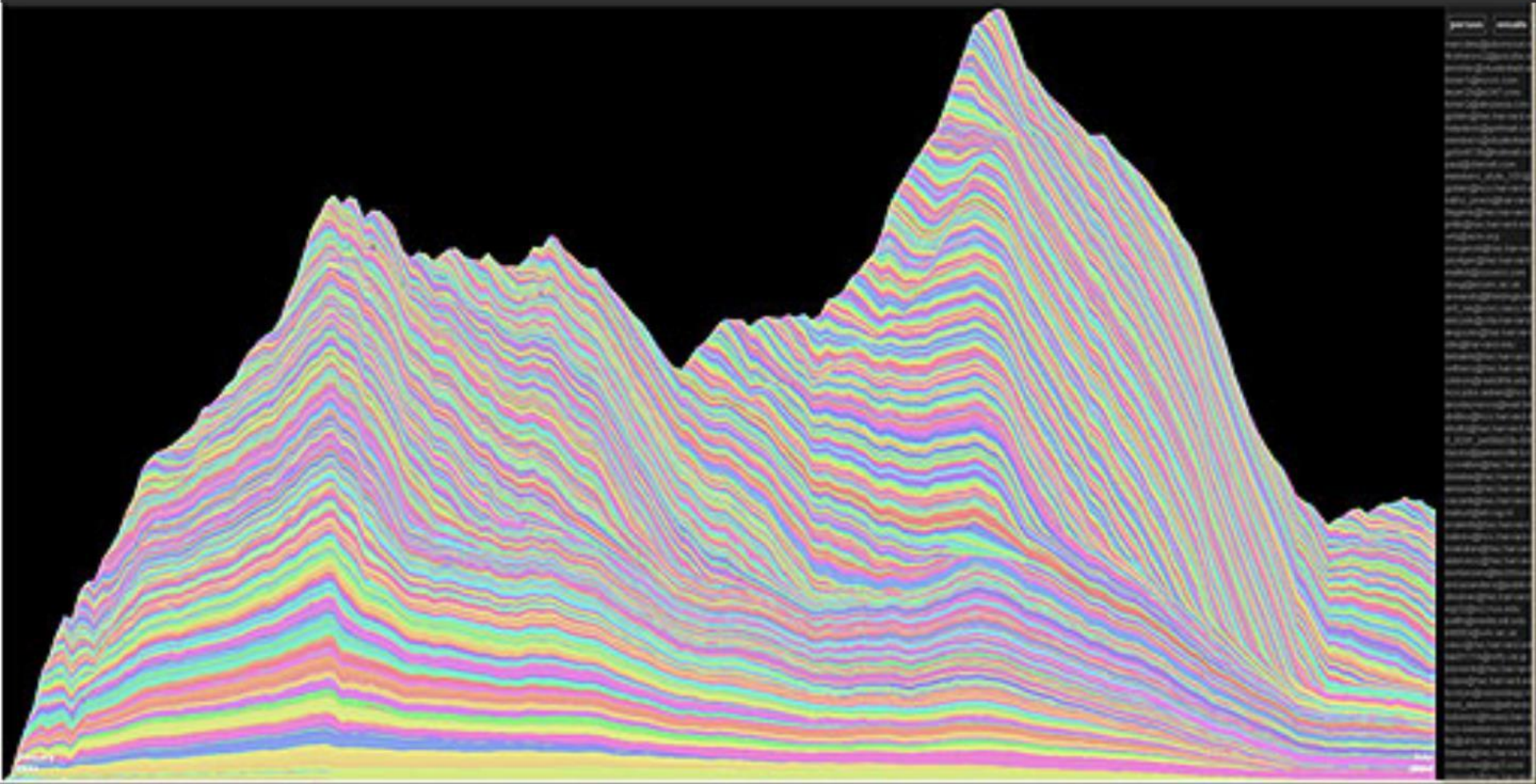
subject	# of posts
Wednesday Spooker ASF	21
WET #3 Anyone for breakfast...	20
Sunny Side Up ASF :)	18
Saturday Ensemble and WET ...	18
Oh no! Watch out! ASF	18
Thursday Combo-Post WET #...	16
The Yellow Rose Inn.....A gift to...	16
WET #1 JBP The First Time ...	16
We Love the Earth ASF	15
Monday Spooker "The Sight" ...	15
C'mon!!!!	14
Theberge "Le Vent Se Leve" ...	14
Holiday Tog #3 :)	13
Spooker du Jour :)	13
Beginning ASF Short and ...	13
Second Try A Kalie for Suzy...	12
Come On a Safari With Me.....	11
Tuesday Spooker ASF	11
Curses, Foiled Again..... ASF ...	10
Halloween Togs Take Two :)	9
Beauty of the Fury Jim Warren...	9
I thought I saw.....? ASF	7
Wednesday Evening at the Con...	4
Second Try A Kalie for Suzy...	2
Frank Was A Monster ASF...	1

subject	# of posts
Sunday Twofer ASF :)	9
Chopsticks/A Jilly fake	8
Oh no! Trouble in Discworld! ...	7
WET..... your thirst ! ASF ...	6
A pretty for you...Reposted fro...	5
Saturday Spooker ASF	5
Sample Previous install Upgr...	4
Tennessee weather tonite	4
WET - Well I am not smiling!	4
Somethin' mushy <asf>	3
Getting seasonal with workin...	3
A Haunted House :)	3
do you wonder what debi's be...	3
Question: Ethics of posters in ...	3
For Jerry	3
Olu's Tribe - slightly rated	3
WET - Glass Bottles	3
Peace Train<ASF>	2
Arrival at Stewart Island !!	2
WET 195 Wrap-up	2
Cat O'Lantern	2
I Put a Spell on You (Happy H...	2
Goodbye to Summer - A Timel...	2
Two Pumpkins In A Strange B...	2
Still Heading South !!	2
WET- Frank Sinatra - The Man...	2
WET Autumn	2
Purple Martin ASF	2
Opposites Attract ...	2
Time	2

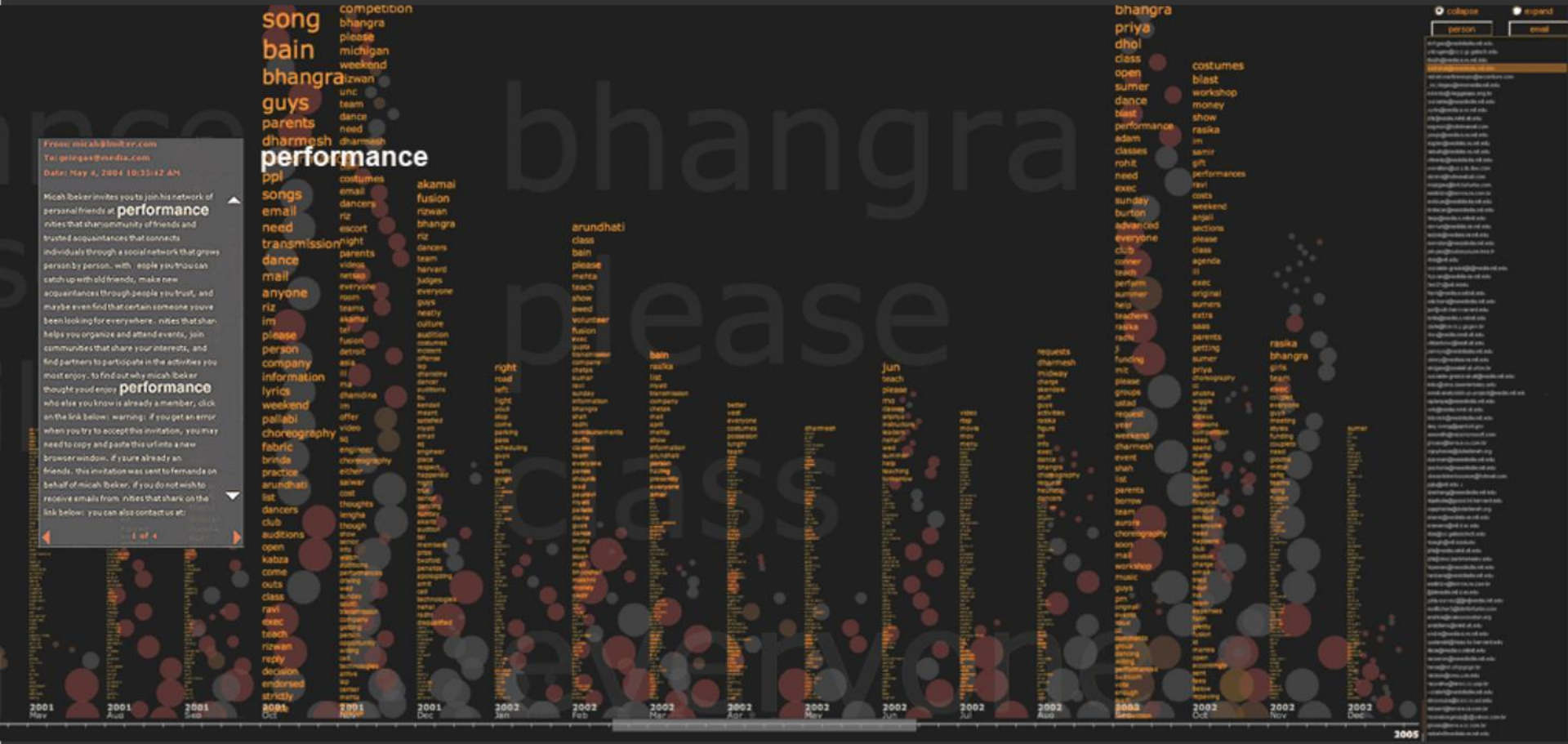

author: rukbat@pern.org
back to newsgroups
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Week of May 6, 2001
Is Genesis Scientifically Correct?
subject
# of posts
threads initiated by author
threads not initiated by author

Email Mountain [Viegas]

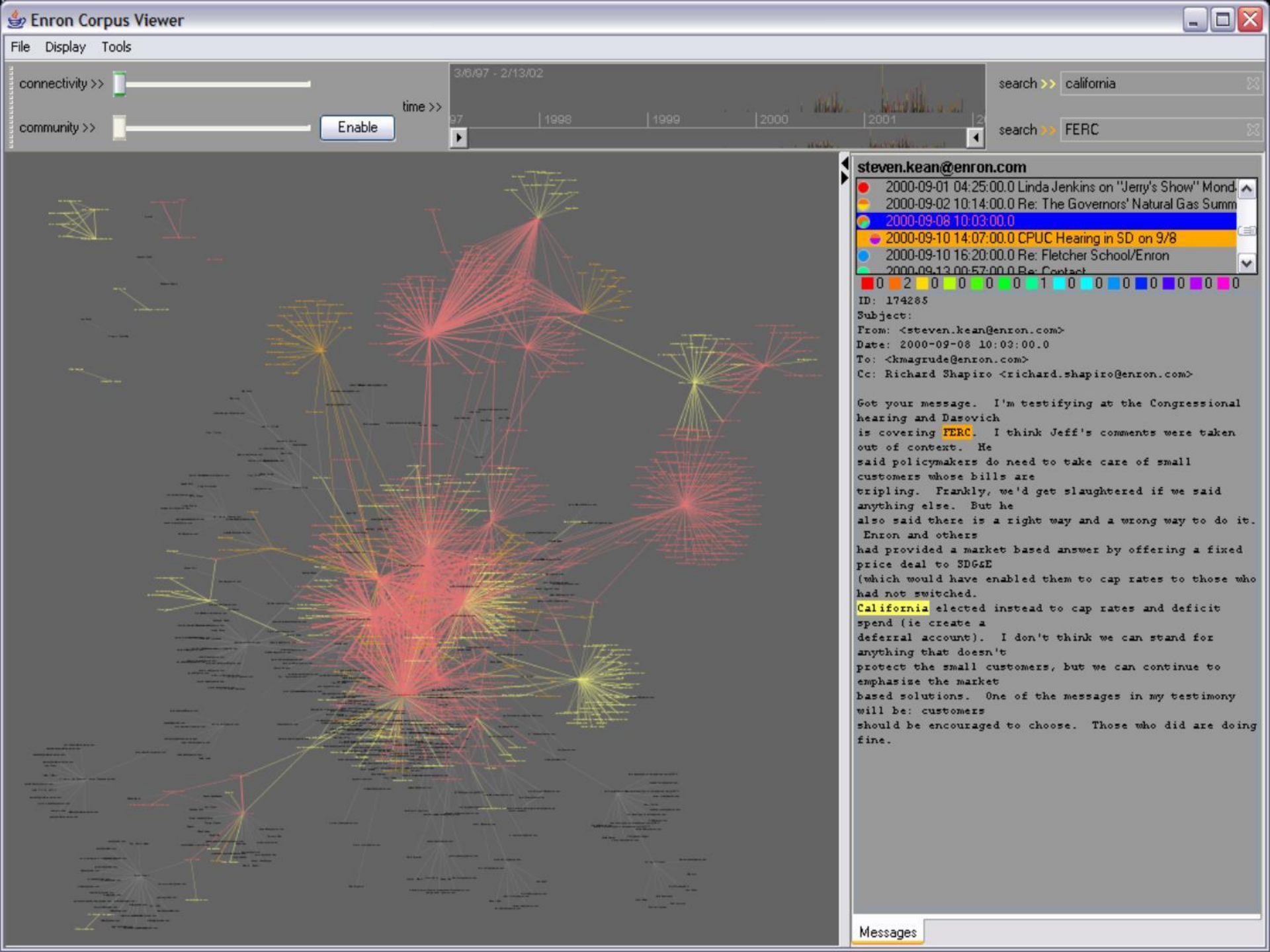
Conversation by person over time (who x when).



Themail [Viegas]



One person over time, TF.IDF weighted terms


Enron Corpus Viewer
File
Display
Tools
connectivity >>
community >>
Enable
time >>
3/6/97 - 2/13/02
1998
1999
2000
2001
search >>
california
search >>
FERC
steven.kean@enron.com
2000-09-01 04:25:00.0 Linda Jenkins on "Jerry's Show" Mond
2000-09-02 10:14:00.0 Re: The Governors' Natural Gas Summ
2000-09-08 10:03:00.0
2000-09-10 14:07:00.0 CPUC Hearing in SD on 9/8
2000-09-10 16:20:00.0 Re: Fletcher School/Enron
0 2 0 0 0 0 0 1 0 0 0 0 0 0 0 0
ID: 174285
Subject:
From: <steven.kean@enron.com>
Date: 2000-09-08 10:03:00.0
To: <kmagrude@enron.com>
Cc: Richard Shapiro <richard.shapiro@enron.com>
Got your message. I'm testifying at the Congressional hearing and Dasovich
is covering FERC. I think Jeff's comments were taken out of context. He
said policymakers do need to take care of small customers whose bills are
tripling. Frankly, we'd get slaughtered if we said anything else. But he
also said there is a right way and a wrong way to do it. Enron and others
had provided a market based answer by offering a fixed price deal to SDG&E
(which would have enabled them to cap rates to those who had not switched.
California elected instead to cap rates and deficit spend (ie create a
deferral account). I don't think we can stand for anything that doesn't
protect the small customers, but we can continue to emphasize the market
based solutions. One of the messages in my testimony will be: customers
should be encouraged to choose. Those who did are doing fine.
Messages

Enron Corpus Viewer

File Display Tools

connectivity >>

community >>

Enable

time >>

1/20/01 - 6/27/01

Feb Mar Apr May Jun

search >> california

search >> ferc

Tim Belden

2001-06-06 16:48:00.0 ISO's Response to BPA Rebuttal of Sh
2001-06-07 11:00:00.0 Legislative Update -- Two Track In The
2001-06-18 00:15:00.0 White House To Support FERC Action
2001-06-19 04:22:00.0 NEWS FLASH ON THIS MORNING'S
2001-06-20 10:37:00.0 Today's Senate Hearing

Washington Lobbyist ?

Jeff Dasovich

Susan Mara

john.shelk@enron.com

dwatkiss@bra

Tim Belden

Tom Hoatson

joe.hartsoe@enron.com

Robert Frank

Alan Co

Linda Robertson

mary.hain@enr

miyung.buster@enron.com

james.steffes@enron.com

paul.kaufman@enron.com

Ray Alvarez

Richard Shapiro

Dan Leff

mark.palmer@enron.com

mona.petrochko@enron.com

ann.schmidt@enron.com

Enron 'Mastermind' Pleads Guilty

SAN FRANCISCO, Oct. 17, 2002

Deputy Attorney General Larry Thompson, center, head of the Justice Departments Corporate Fraud Task Force, comments Thursday on the guilty plea by Timothy N. Belden, Enron's chief energy trader. (Photo: CBS/AP)

(AP) A former top energy trader, considered the mastermind of Enron Corp.'s scheme to drive up California's energy prices, pleaded guilty Thursday to a federal conspiracy charge.

Timothy Belden, the former head of trading in Enron's Portland, Ore., office, admitted to one count of conspiracy to commit wire fraud and promised to cooperate with state and federal prosecutors as well as any non-criminal effort to investigate the energy industry.

"I did it because I was trying to maximize profit for Enron," Belden told U.S. District Judge Martin Jenkins.

that the FERC order
ing from the western
), for refunds
n essence, they are
ew order applied to
e order.
flavor is that Sen.
for deregulatioin
(by this point the
order applied only
t tried to
that Con Ed has
York and he could
wski said he had

from four western governors -- those from Arizona, North Dakota, Utah and
Wyoming -- saying that since FERC has acted, there is no need for Congress to
pursue price control legislation.

There were a series of questions and comments on details and technical
aspects of the orders. I will do an e-mail on these items later today.
Please advise if you have any questions or comments.

Messages

Document Collections

Named Entity Recognition

Label named entities in text:

John Smith -> PERSON

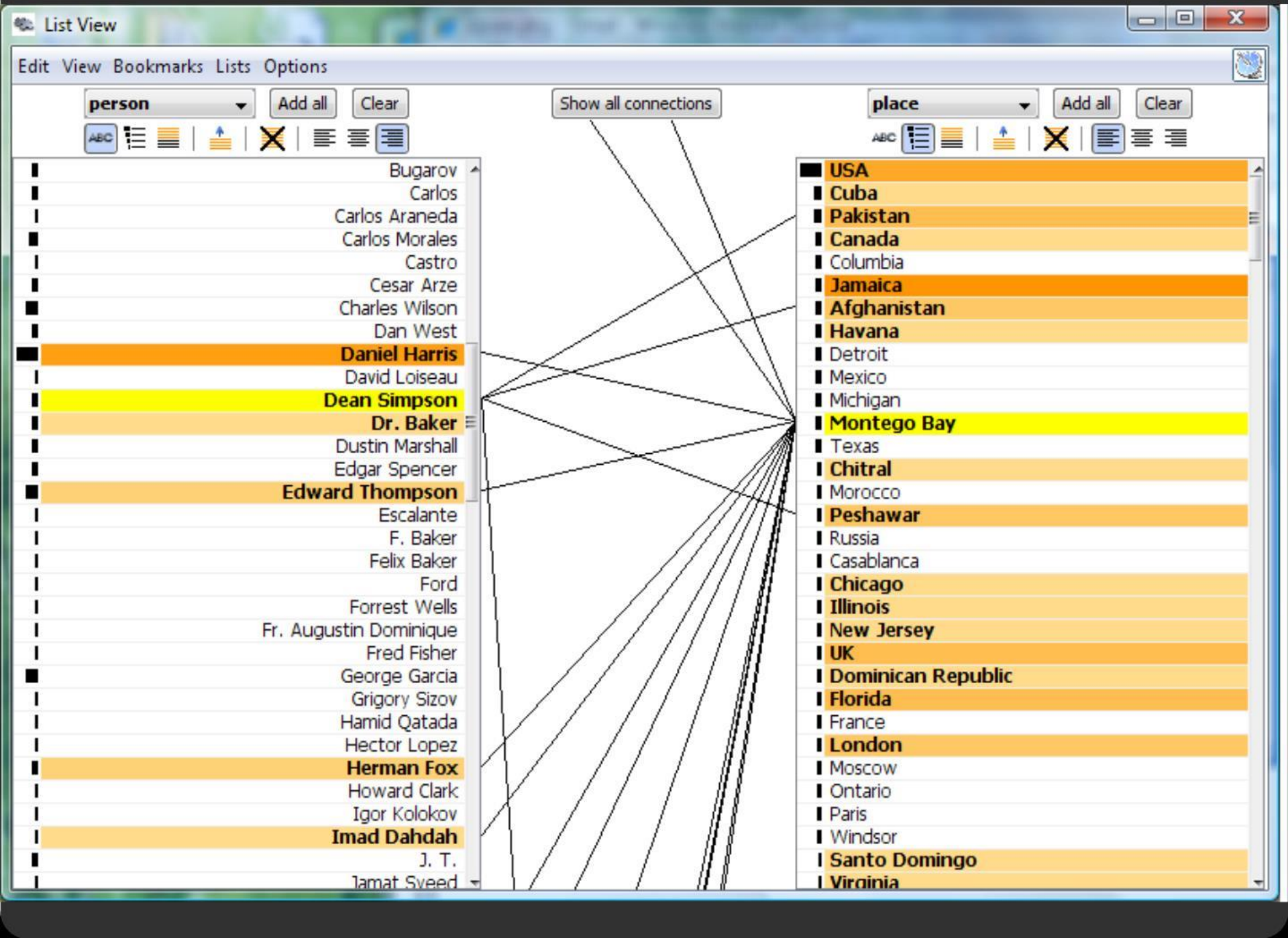
Soviet Union -> COUNTRY

353 Serra St -> ADDRESS

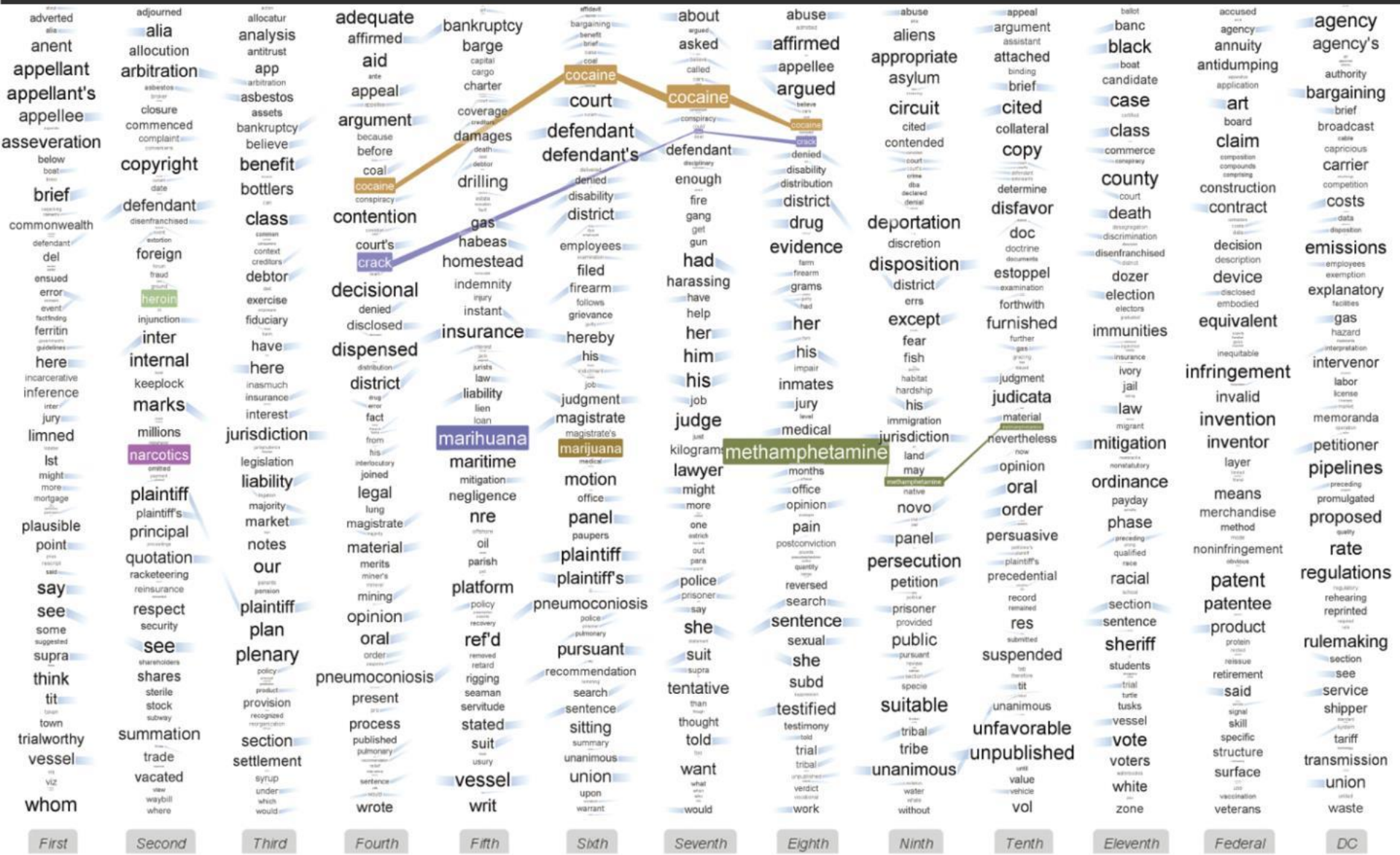
(555) 721-4312 -> PHONE NUMBER

Entity relations: how do the entities relate?

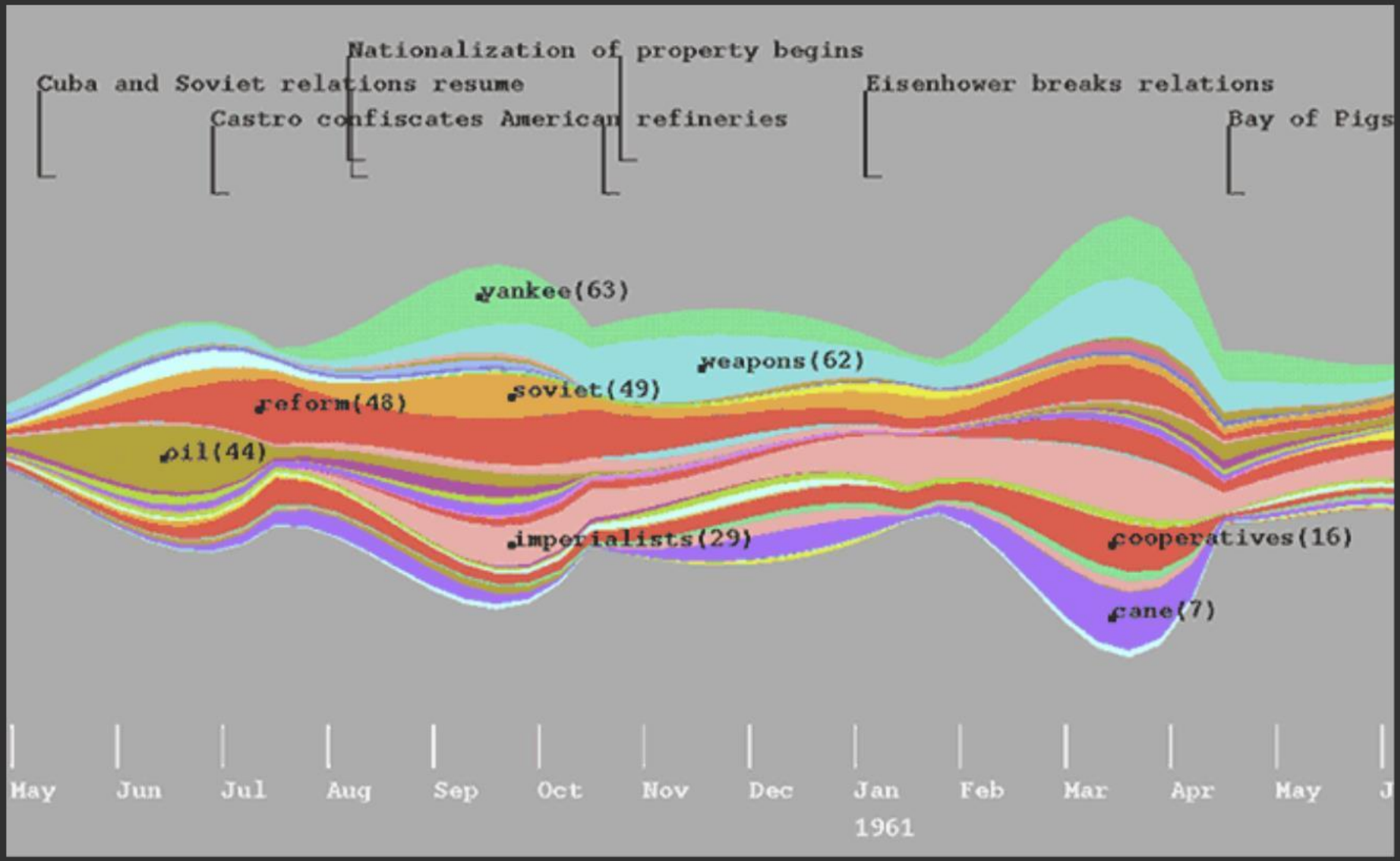
Simple approach: do they co-occur in a small window of text?


List View
Edit View Bookmarks Lists Options
person
Add all
Clear
Show all connections
place
Add all
Clear
Bugarov
Carlos
Carlos Araneda
Carlos Morales
Castro
Cesar Arze
Charles Wilson
Dan West
Daniel Harris
David Loiseau
Dean Simpson
Dr. Baker
Dustin Marshall
Edgar Spencer
Edward Thompson
Escalante
F. Baker
Felix Baker
Ford
Forrest Wells
Fr. Augustin Dominique
Fred Fisher
George Garcia
Grigory Sizov
Hamid Qatada
Hector Lopez
Herman Fox
Howard Clark
Igor Kolokov
Imad Dahdah
J. T.
Jamat Syeed
USA
Cuba
Pakistan
Canada
Columbia
Jamaica
Afghanistan
Havana
Detroit
Mexico
Michigan
Montego Bay
Texas
Chitral
Morocco
Peshawar
Russia
Casablanca
Chicago
Illinois
New Jersey
UK
Dominican Republic
Florida
France
London
Moscow
Ontario
Paris
Windsor
Santo Domingo
Virginia

Parallel Tag Clouds [Collins et al.]



Theme River [Havre et al.]



Similarity & Clustering

Compute vector distance among docs

For TF.IDF, typically cosine distance

Similarity measure can be used to cluster

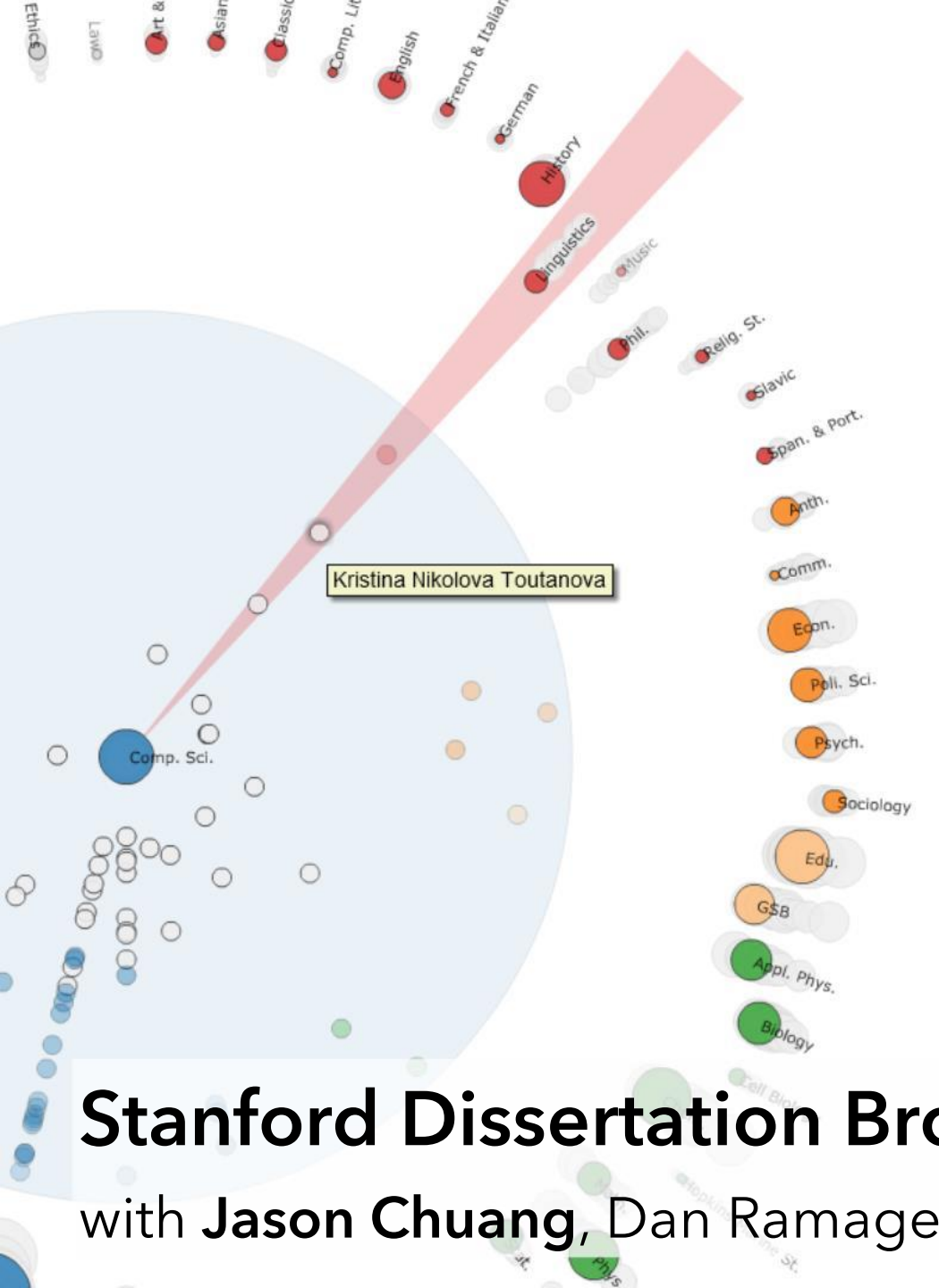
Topic modeling

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



Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

Computer Science (2005)

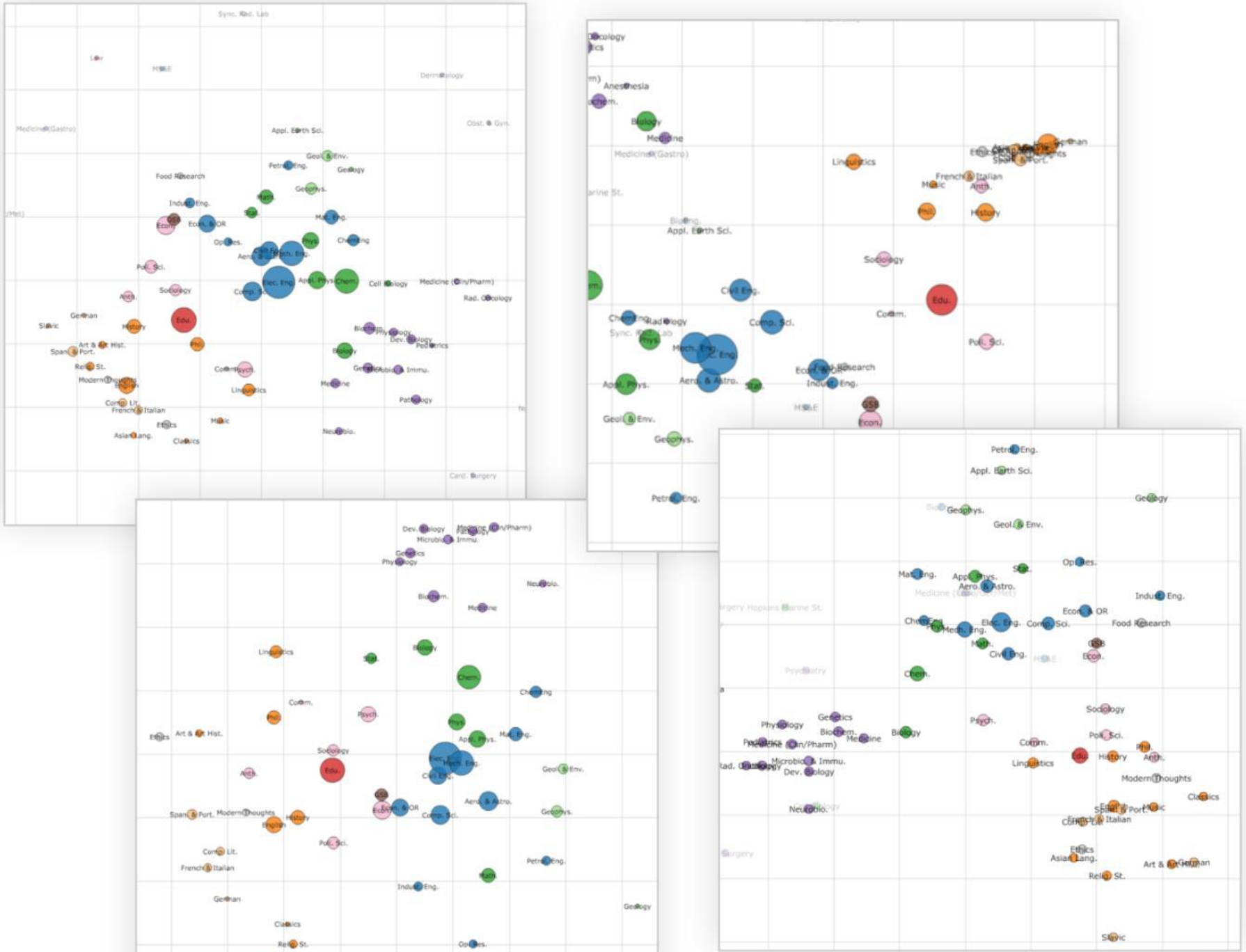
Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks--sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.

Stanford Dissertation Browser

with **Jason Chuang**, Dan Ramage & Christopher Manning


Sync. Rad. Lab
Dermatology
Medicine (Gastro)
Obst. & Gyn.
Appl. Earth Sci.
Geol. & Env.
Petrol. Eng.
Geology
Food Research
Geophys.
Indust. Eng.
Stat.
Econ.
Econ. & OR
Mat. Eng.
Optics
Phys.
Chem. Eng.
Aero. & Astro.
Mech. Eng.
Elec. Eng.
Appl. Phys.
Chem.
Cell Biology
Medicine (Clin/Pharm)
Rad. Oncology
Pol. Sci.
Sociology
Comp. Sci.
Civil Eng.
Anesthesia
Biology
Medicine
Linguistics
French & Italian
Music
Anth.
Phil.
History
Edu.
Comm.
Slavic
German
Art & Art Hist.
Span. & Port.
Relig. St.
Modern Thoughts
English
Comp. Lit.
Ethics
Asian Lang.
Classics
Psych.
Biochem.
Genetics
Physiology
Microbiol. & Immu.
Dev. Biology
Pediatrics
Neurobio.
Pathology
Care. Surgery

Thank you