Text Visualization

Maneesh Agrawala

CS 448B: Visualization
Fall 2020

1

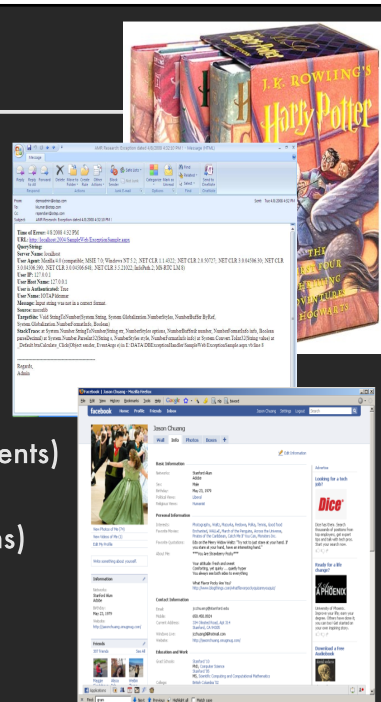
Text as data

Documents

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

Collection of documents

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



2

Announcements

3

Final project

Data analysis/explainer or conduct research

- **Data analysis**: Analyze dataset in depth & make a visual explainer
- **Research**: Pose problem, Implement creative solution

Deliverables

- **Data analysis/explainer**: Article with multiple interactive visualizations
- **Research**: Implementation of solution and web-based demo if possible
- **Short video (2 min)** demoing and explaining the project

Schedule

- Project proposal: Thu 10/29
- Design Review and Feedback: Tue 11/17 & Thu 11/19
- Final code and video: Sat 11/21 11:59pm

Grading

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

4

Class Schedule

Guest Lecture Th Nov 12

Jessica Hullman on
Visualizing Uncertainty

5

Design Feedback (Next Week)

Signup for a ~10 min slot

Will post signups on Piazza later this week

Plan to give a 5 min presentation (mostly demo) of work so far. We will give oral feedback.

6

Text Visualization

8

Why visualize text?

9

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

10

Example: Health Care Reform

Background

- 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?

11

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 you

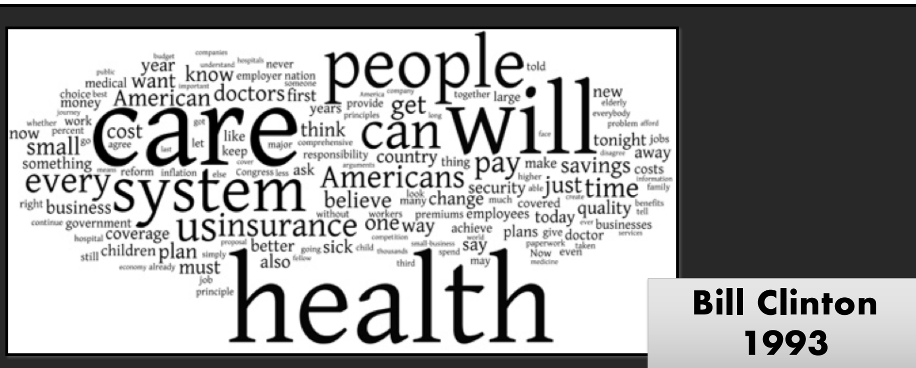
12

Word/Tag Clouds: Word Count

President Obama's Health Care Speech to Congress

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

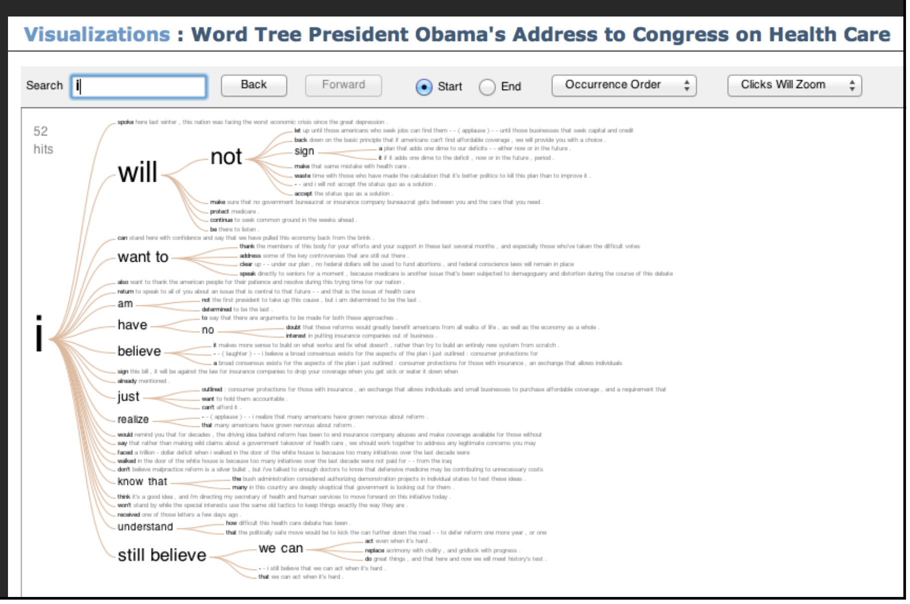
13



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

14

WordTree: Word Sequences



15

16

Gulf of Evaluation

Many (most?) text visualizations do not represent 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?

17

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

18

Topics

Text as Data

Visualizing Document Content

Visualizing Conversation

Document Collections

20

Text as Data

21

Words as nominal data?

High dimensional (10,000+)

More than equality tests

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

Words have meanings and relations

22

Text Processing Pipeline

Tokenization

Segment text into terms.
Remove stop words? *a, an, the, of, to, be*
Numbers and symbols? *#cardinal, @Stanford, OMG!!!!!!!!*
Entities? *Palo Alto, O'Connor, U.S.A.*

23

Text Processing Pipeline

Tokenization

Segment text into terms.
Remove stop words? *a, an, the, of, to, be*
Numbers and symbols? *#cardinal, @Stanford, OMG!!!!!!!!*
Entities? *Palo Alto, O'Connor, U.S.A.*

Stemming

Group together different forms of a word.
Porter stemmer? *visualization(s), visualize(s), visually* -> *visual*
Lemmatization? *goes, went, gone* -> *go*

24

Text Processing Pipeline

Tokenization

Segment text into terms.
Remove stop words? *a, an, the, of, to, be*
Numbers and symbols? *#cardinal, @Stanford, OMG!!!!!!!!*
Entities? *Palo Alto, O'Connor, U.S.A.*

Stemming

Group together different forms of a word.
Porter stemmer? *visualization(s), visualize(s), visually* -> *visual*
Lemmatization? *goes, went, gone* -> *go*

Ordered list of terms

25

The Bag of Words Model

Ignore ordering relationships within the text

A document ≈ vector of term weights

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

Aggregate into a document x term matrix

Document vector space model

26

Document x Term matrix

Each document is a vector of term weights
Simplest weighting is to just count occurrences

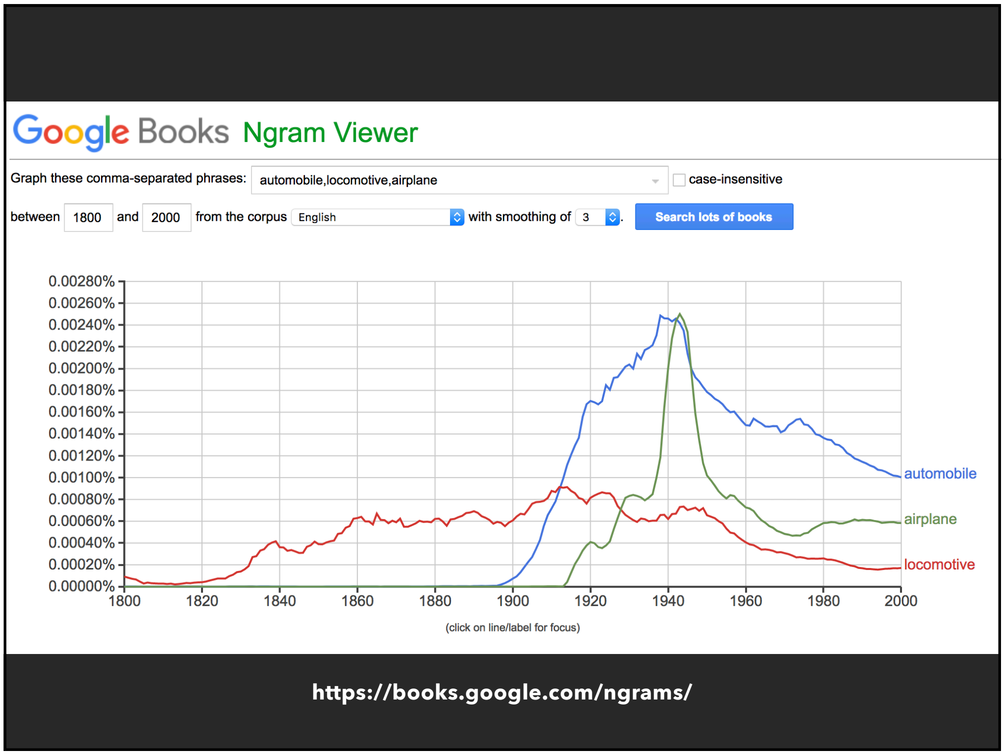
	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

27

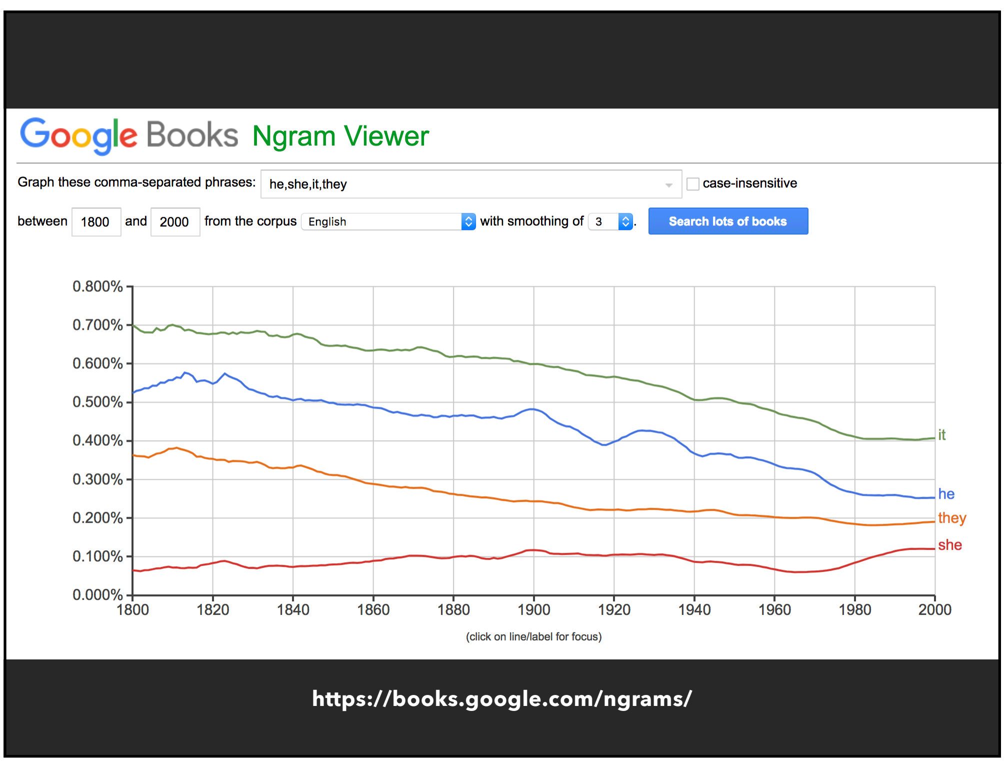
WordCount (Harris 2004)

http://wordcount.org

28


Google Books Ngram Viewer
Graph these comma-separated phrases: automobile,locomotive,airplane
case-insensitive
between 1800 and 2000 from the corpus English with smoothing of 3.
Search lots of books
0.00280%
0.00260%
0.00240%
0.00220%
0.00200%
0.00180%
0.00160%
0.00140%
0.00120%
0.00100%
0.00080%
0.00060%
0.00040%
0.00020%
0.00000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
automobile
airplane
locomotive
(click on line/label for focus)
https://books.google.com/ngrams/

29


Google Books Ngram Viewer
Graph these comma-separated phrases: he,she,it,they
case-insensitive
between 1800 and 2000 from the corpus English with smoothing of 3.
Search lots of books
0.800%
0.700%
0.600%
0.500%
0.400%
0.300%
0.200%
0.100%
0.000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
it
he
they
she
(click on line/label for focus)
https://books.google.com/ngrams/

30

31

Word/Tag Clouds

Strengths

Can help with gisting and initial query formation

Weaknesses

Sub-optimal visual encoding (size not pos. encodes freq.)
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

32

Given a text, what are the best descriptive words?

33

Keyword Weighting

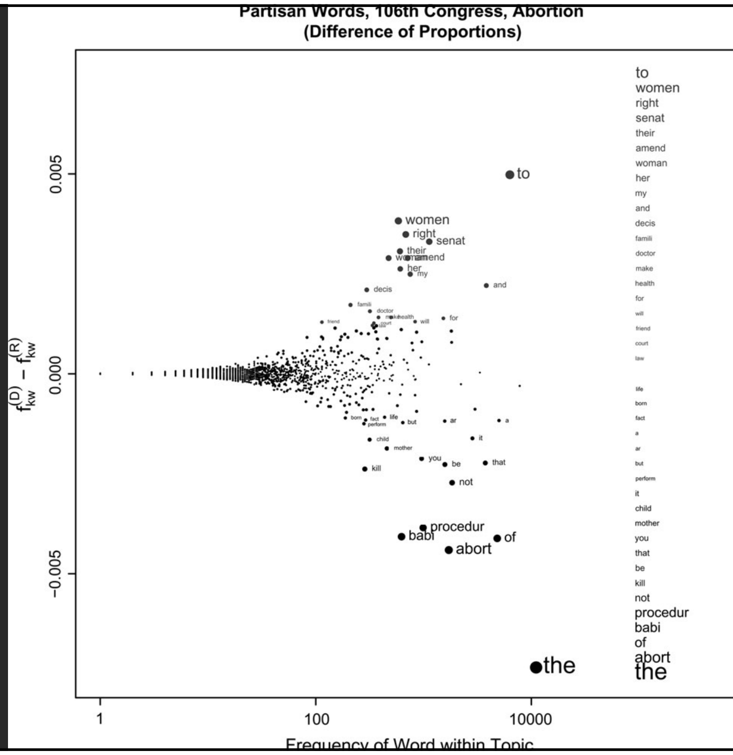
Term Frequency

tf_{td} = count(t) in d

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

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

34



35

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

36

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

"Bag of words" ignores additional info.

Grammar / part-of-speech

Position within document

Recognizable entities

42

How do people describe text?

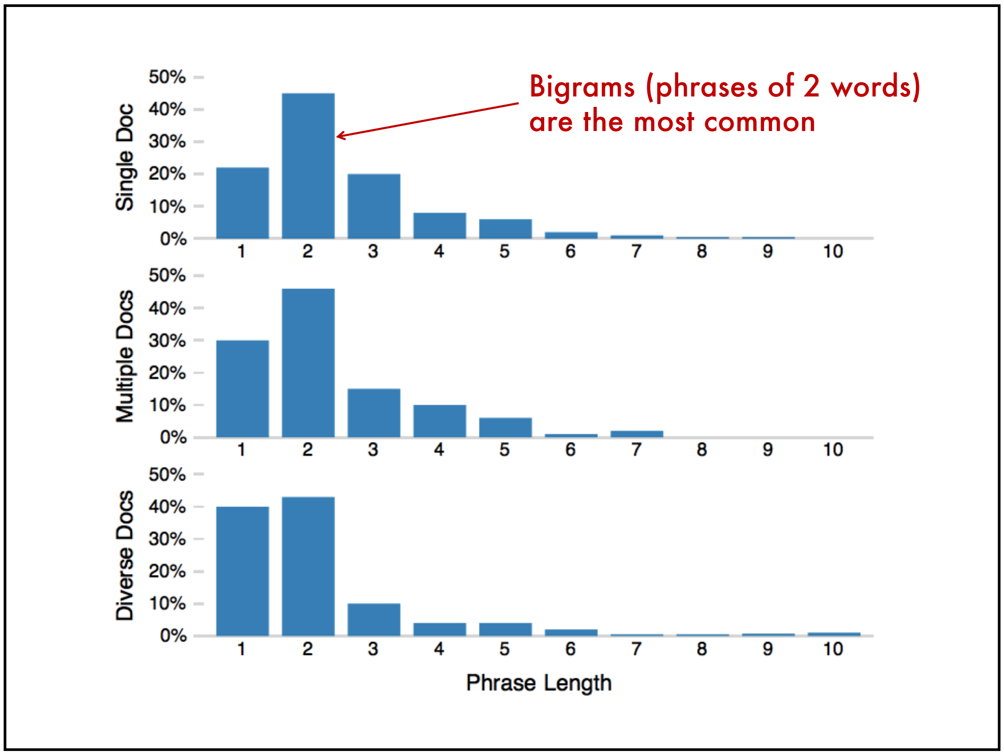
Asked 69 graduate students to read and describe dissertation abstracts

Each given 3 documents in sequence; summarized each using keyphrases, then summarized the 3 together as a whole using keyphrases

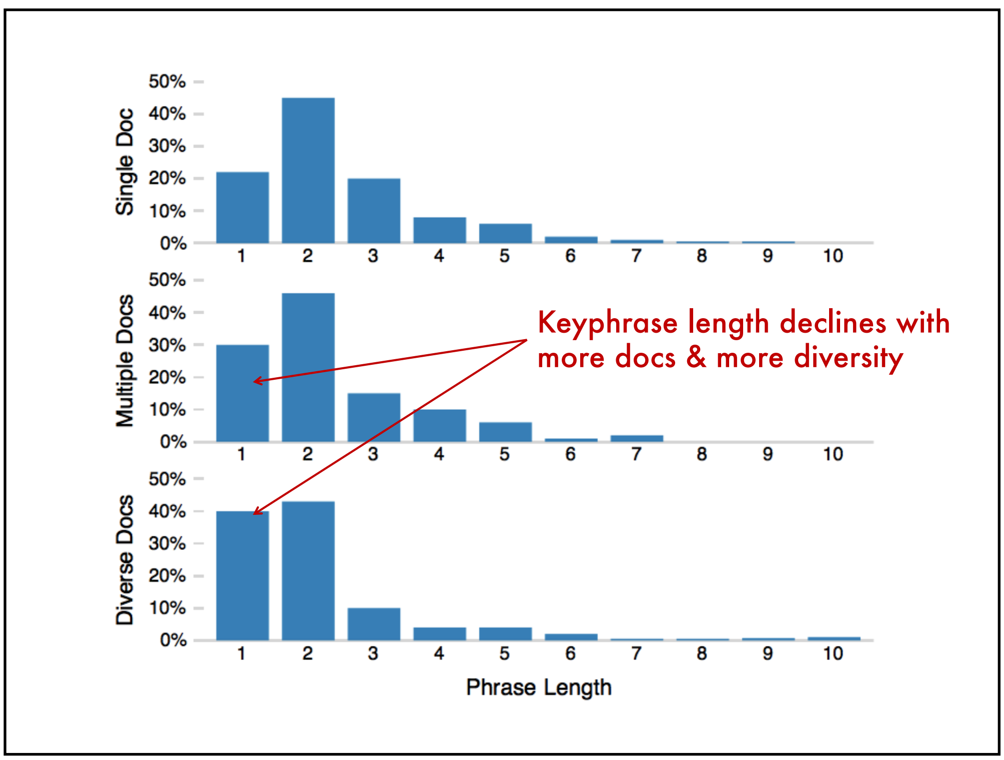
Were matched to both *familiar* and *unfamiliar* topics; *topical diversity* within a collection was varied systematically

[Chuang 2012]

43


Bigrams (phrases of 2 words) are the most common
Single Doc
Multiple Docs
Diverse Docs
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10
Phrase Length

44


Keyphrase length declines with more docs & more diversity
Single Doc
Multiple Docs
Diverse Docs
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10
Phrase Length

45

Term Commonness

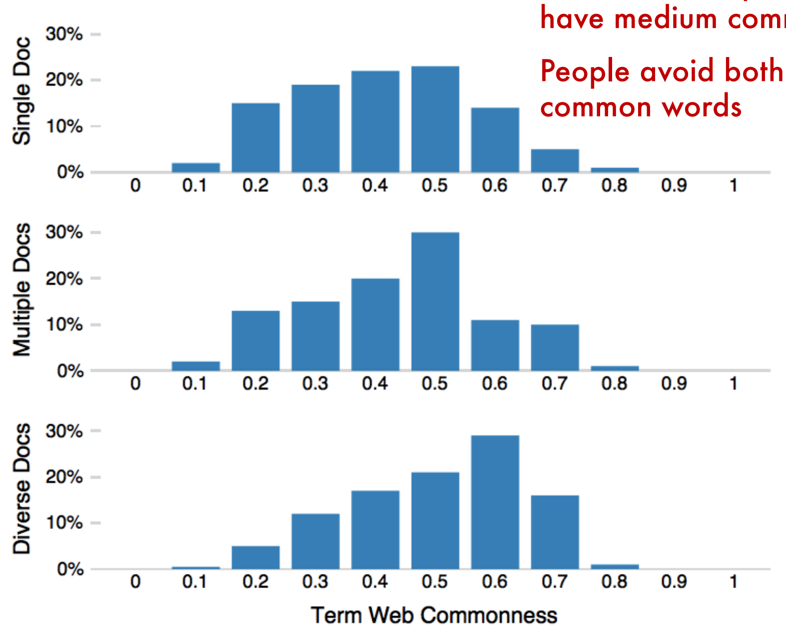
$$\log(tf_w) \;/\; \log(tf_{the})$$

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

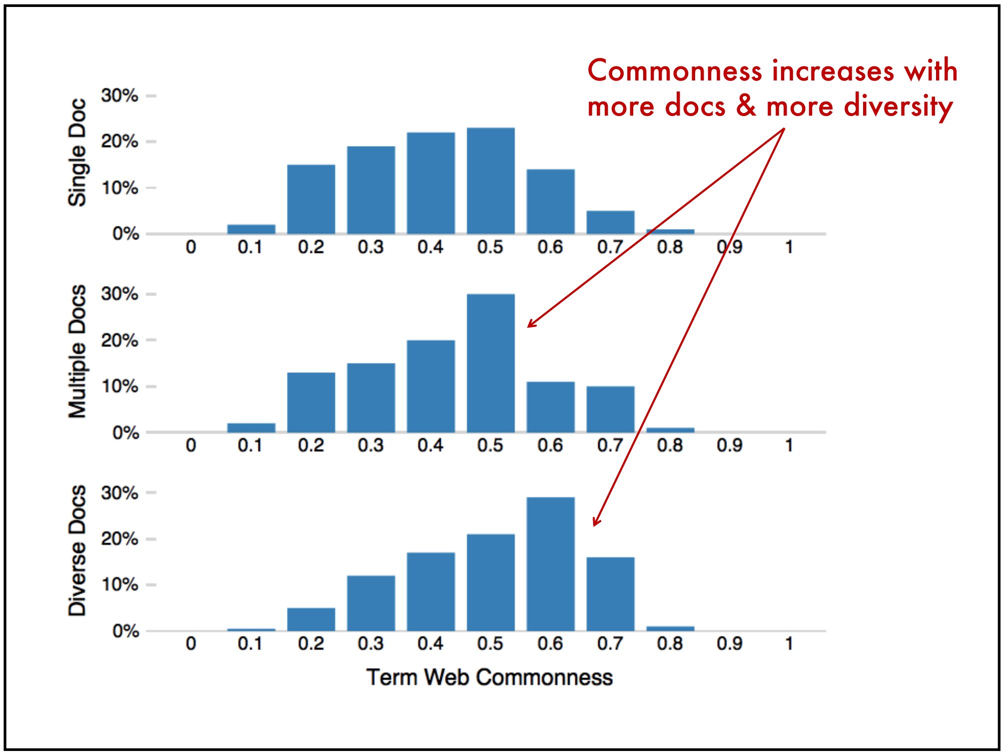
46

Selected descriptive terms have medium commonness

People avoid both rare and common words



47

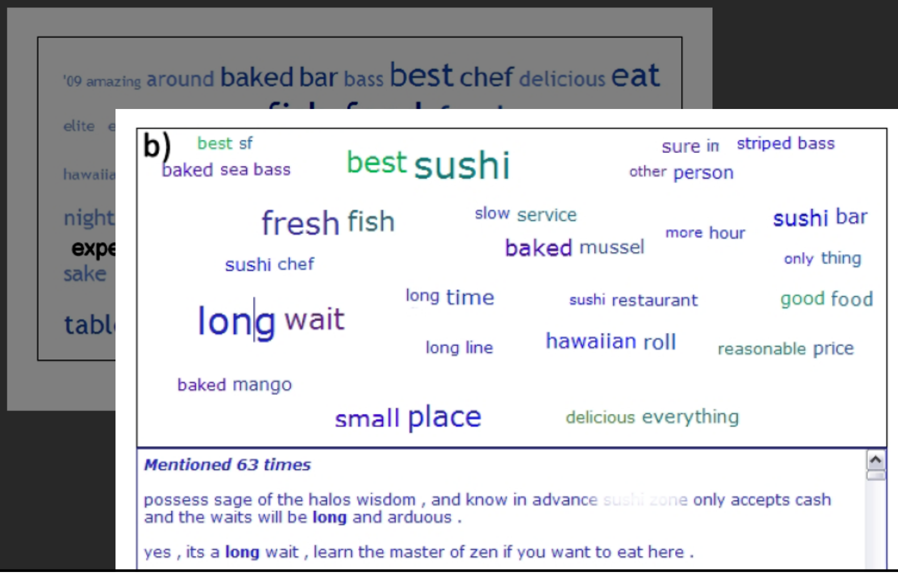

Commonness increases with more docs & more diversity
Single Doc
Multiple Docs
Diverse Docs
30%
20%
10%
0%
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Term Web Commonness

48


Yelp: Review Spotlight [Yatani 2011]
'09 amazing around baked bar bass best chef delicious eat
elite everything favorite fish food fresh going hamachi
hawaiian hour line love mango minutes mussels name
night nigiri order people prices really restaurant roll
expensive or cheap?
sake salmon sea seated service spicy stars sure sushi
table think tuna wait waitress worth
"long wait" or "no wait"?
what type of sushi roll?

52

Yelp: Review Spotlight [Yatani 2011]



53

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

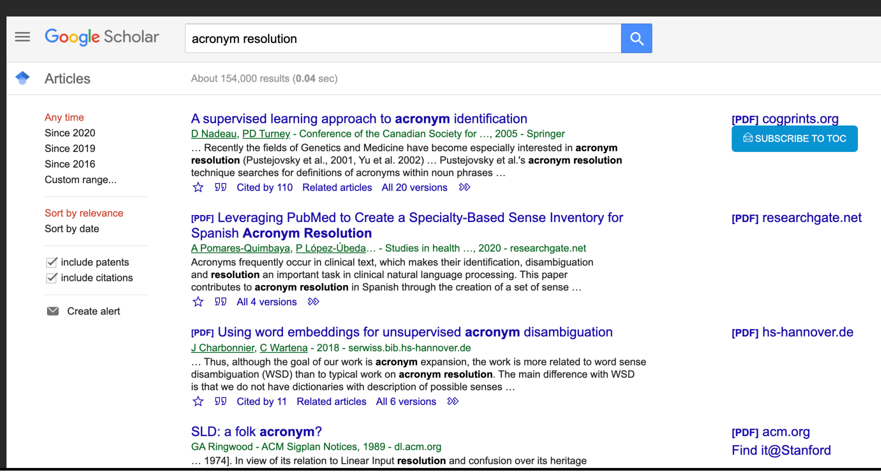
54

Visualizing Document Content

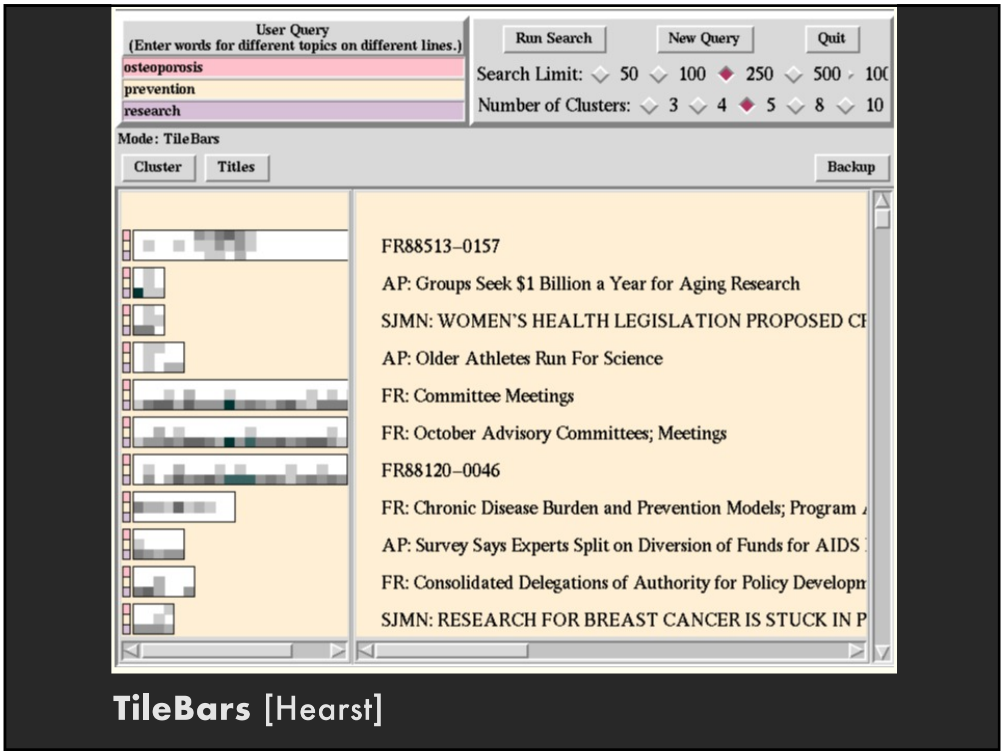
55

Information Retrieval

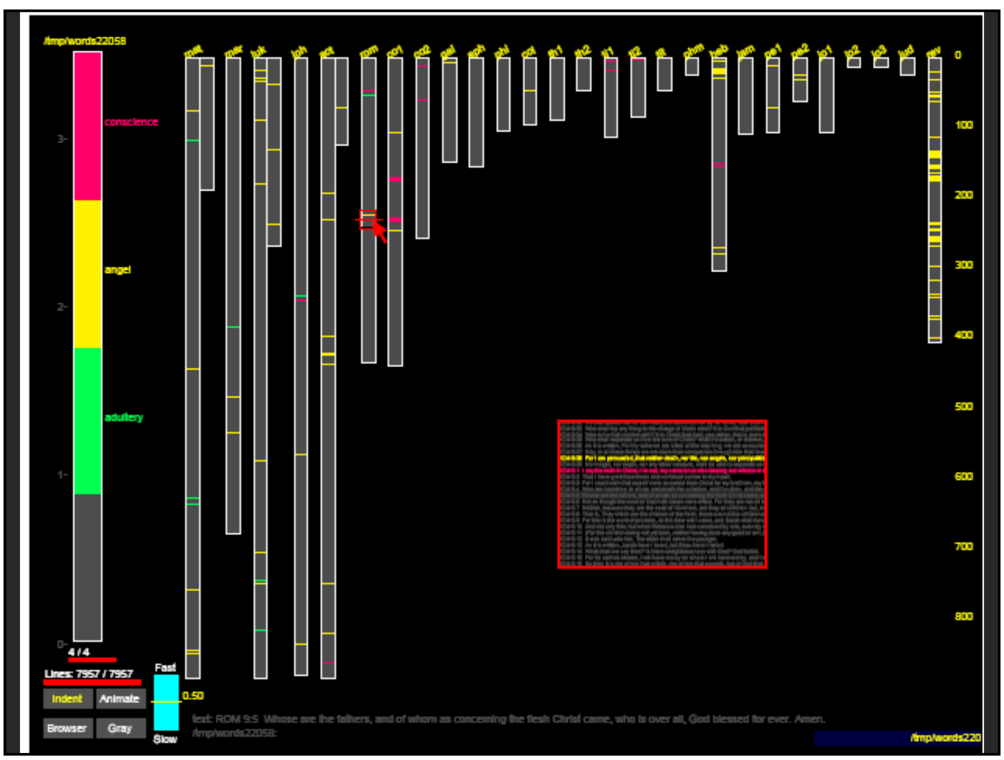
Search for documents
Match query string with documents
Visualization to **contextualize results**



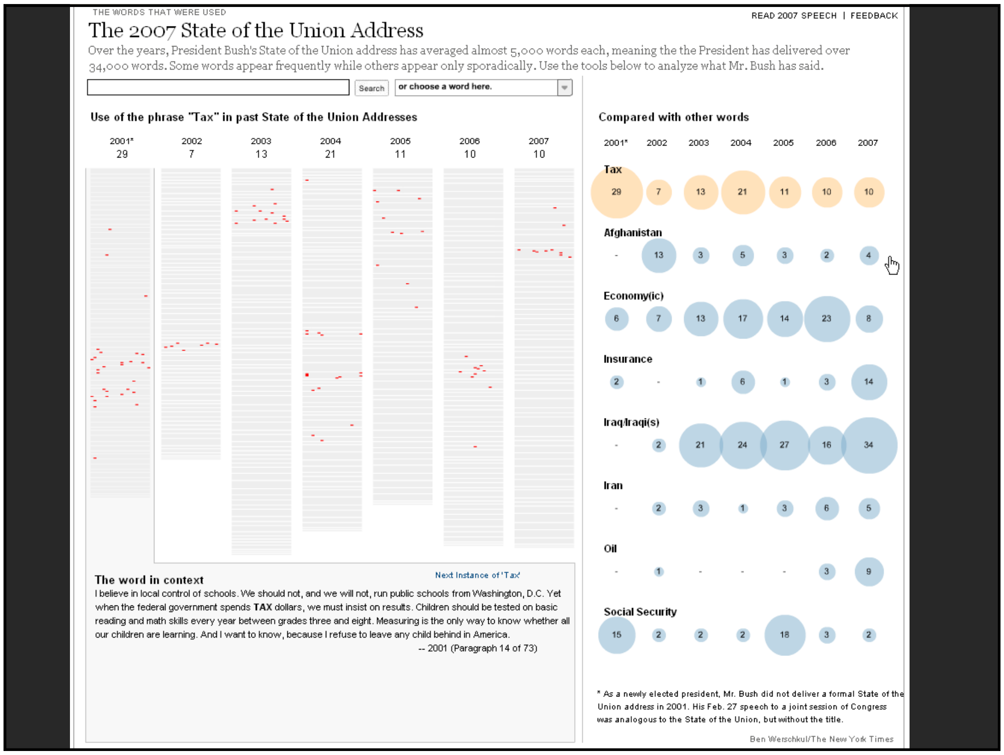
56


User Query
(Enter words for different topics on different lines.)
osteoporosis
prevention
research
Run Search
New Query
Quit
Search Limit: 50 100 250 500 10(
Number of Clusters: 3 4 5 8 10
Mode: TileBars
Cluster
Titles
Backup
FR88513–0157
AP: Groups Seek $1 Billion a Year for Aging Research
SJMN: WOMEN'S HEALTH LEGISLATION PROPOSED CI
AP: Older Athletes Run For Science
FR: Committee Meetings
FR: October Advisory Committees; Meetings
FR88120–0046
FR: Chronic Disease Burden and Prevention Models; Program
AP: Survey Says Experts Split on Diversion of Funds for AIDS
FR: Consolidated Delegations of Authority for Policy Developn
SJMN: RESEARCH FOR BREAST CANCER IS STUCK IN P
TileBars [Hearst]

57


/tmp/words22058
conscience
angel
adultery
4 / 4
Lines: 7957 / 7957
Indent
Animate
Browser
Gray
Fast
0.50
Slow
0
100
200
300
400
500
600
700
800

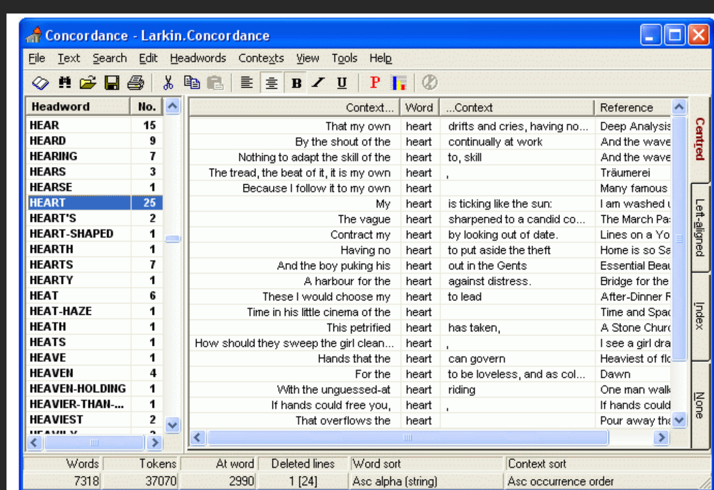
58



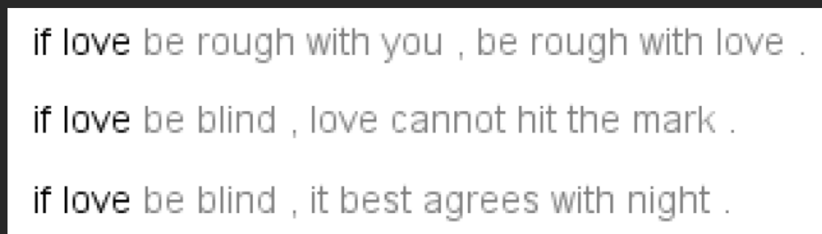
59

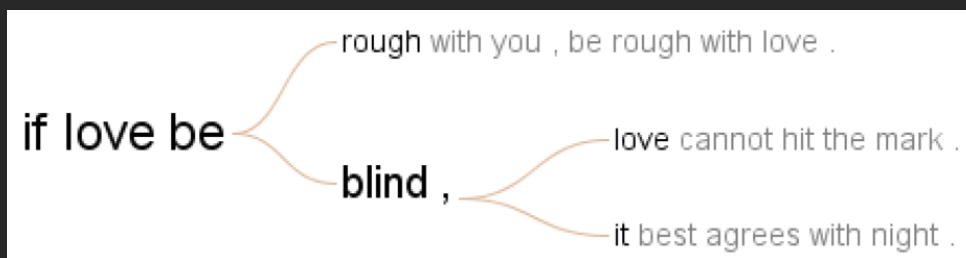
Concordance

What is the common local context of a term?



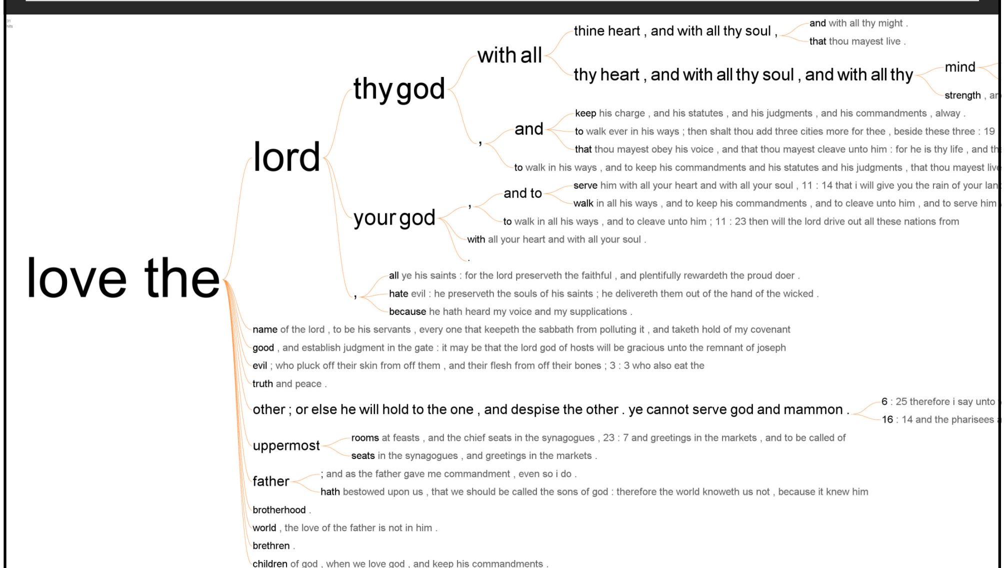
62





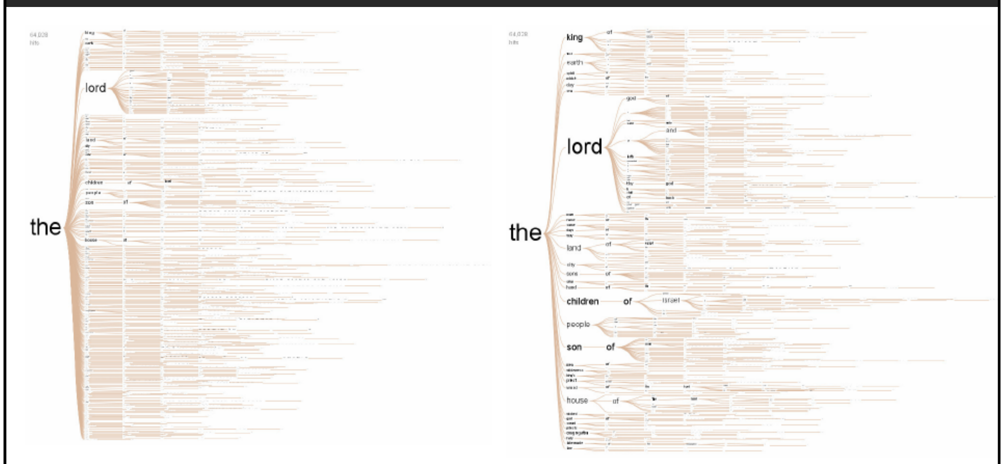
64

WordTree



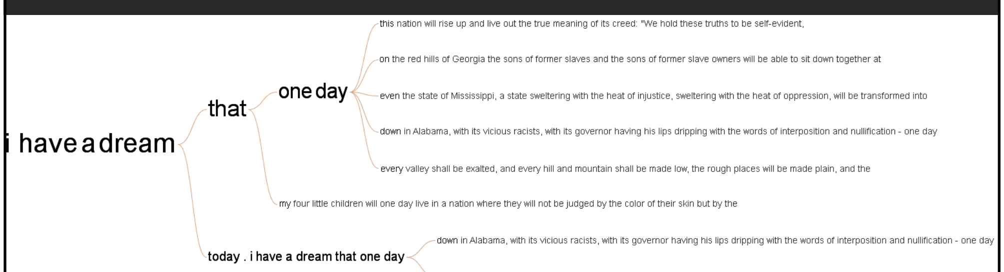
65

Filter infrequent runs

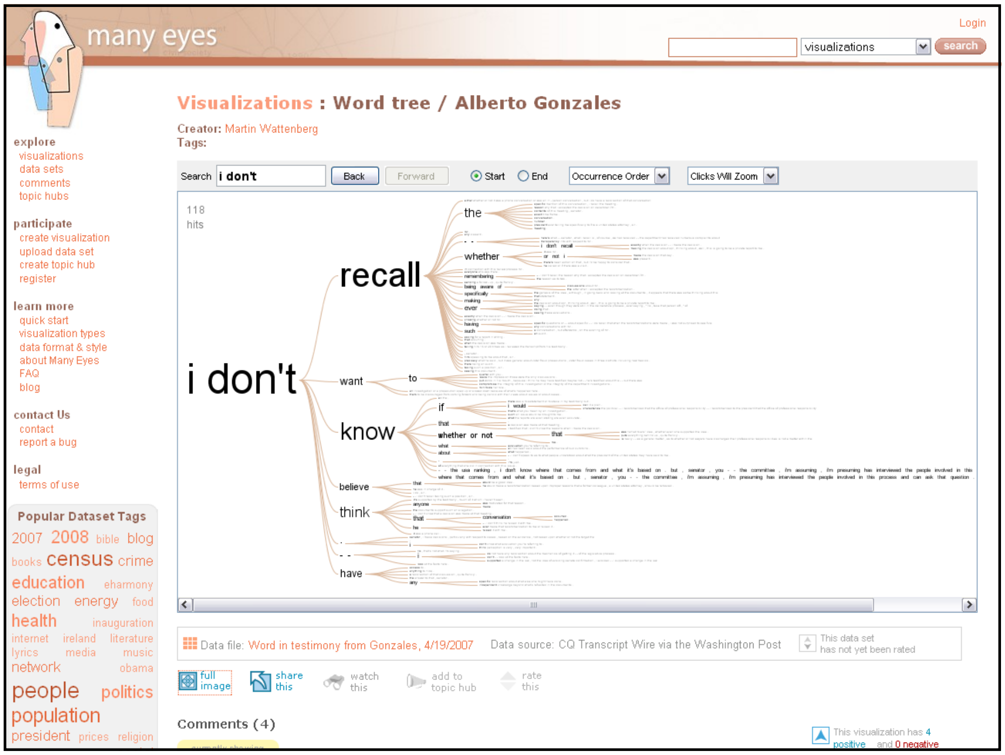


66

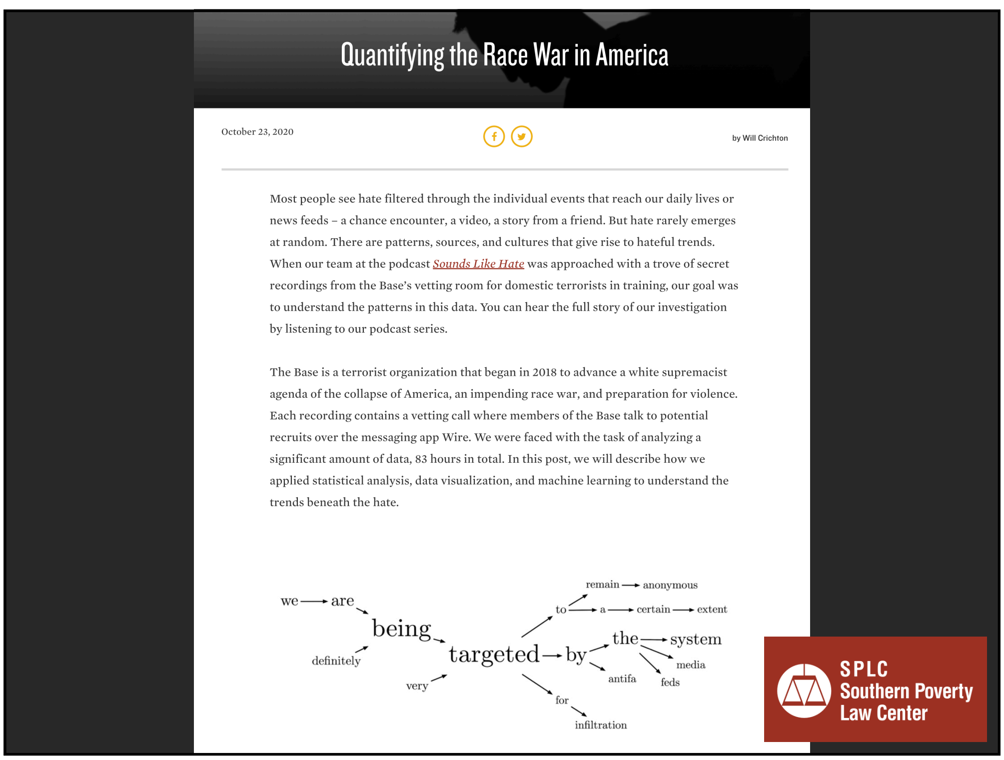
Recurrent themes in speech



67


many eyes
Login
visualizations
search
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
118 hits
i don't
recall
the
whether
want
to
know
if
believe
think
have
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)
This visualization has 4 positive and 0 negative

68


Quantifying the Race War in America
October 23, 2020
by Will Crichton
Most people see hate filtered through the individual events that reach our daily lives or news feeds – a chance encounter, a video, a story from a friend. But hate rarely emerges at random. There are patterns, sources, and cultures that give rise to hateful trends. When our team at the podcast Sounds Like Hate was approached with a trove of secret recordings from the Base's vetting room for domestic terrorists in training, our goal was to understand the patterns in this data. You can hear the full story of our investigation by listening to our podcast series.
The Base is a terrorist organization that began in 2018 to advance a white supremacist agenda of the collapse of America, an impending race war, and preparation for violence. Each recording contains a vetting call where members of the Base talk to potential recruits over the messaging app Wire. We were faced with the task of analyzing a significant amount of data, 83 hours in total. In this post, we will describe how we applied statistical analysis, data visualization, and machine learning to understand the trends beneath the hate.
we
are
being
definitely
very
targeted
by
to
remain
anonymous
a
certain
extent
the
system
media
feds
antifa
for
infiltration
SPLC
Southern Poverty
Law Center

69

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>

70

Phrase Nets [van Ham 2009]

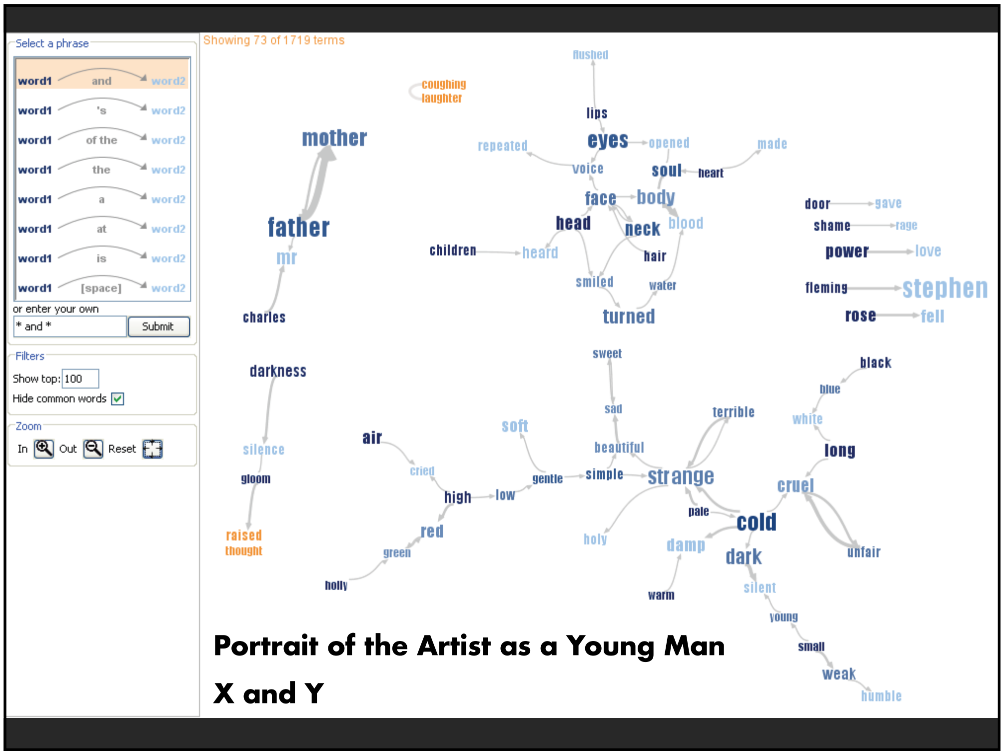
Look for specific linking patterns in the text:

‘A and B’, ‘A at B’, ‘A of B’, etc
Could be output of regexp or parser

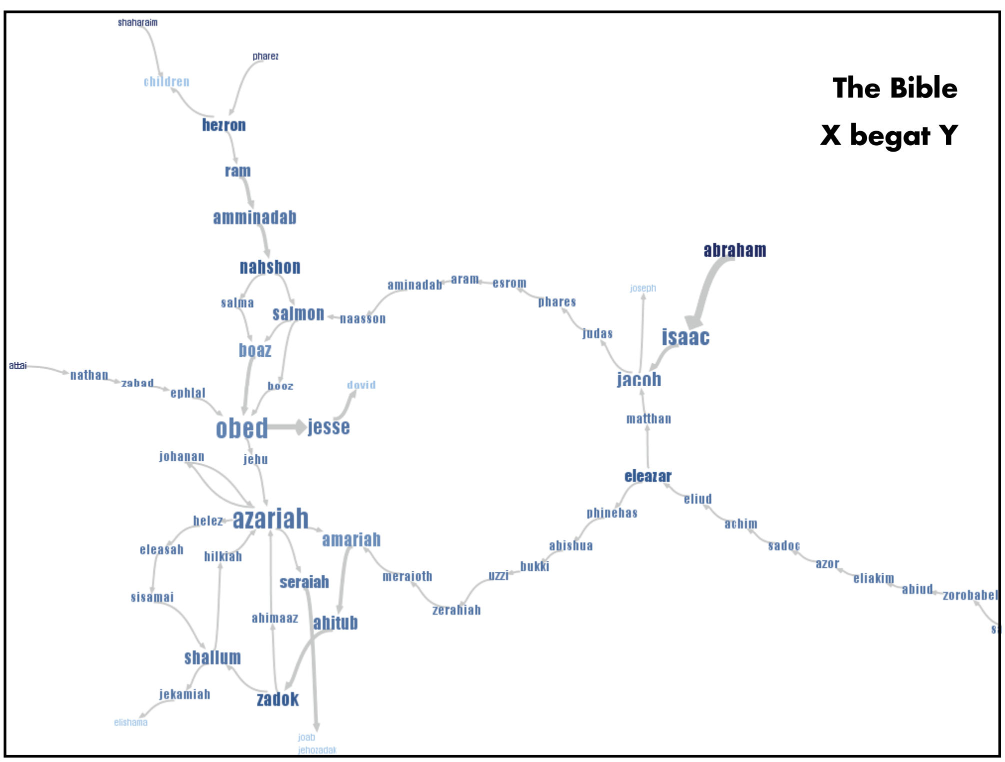
Visualize extracted patterns in a node-link view

Occurrences → Node size
Pattern position → Edge direction
Darker color → higher ratio of out-edges to in-edges

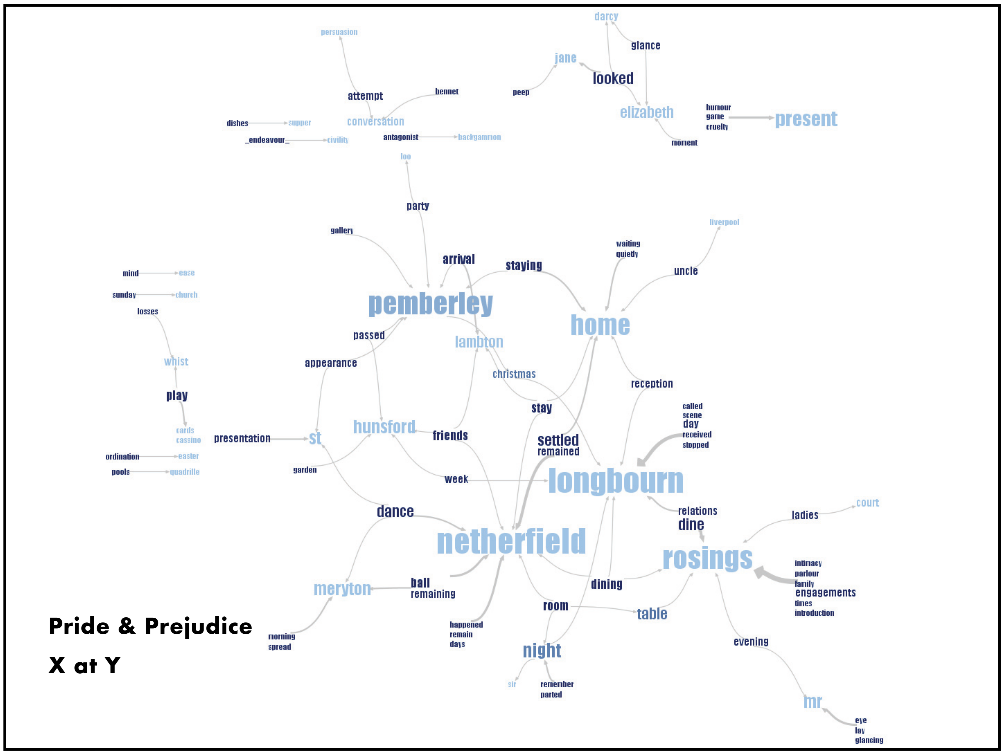
71


Select a phrase
word1 and word2
word1 's word2
word1 of the word2
word1 the word2
word1 a word2
word1 at word2
word1 is word2
word1 [space] word2
or enter your own
* and *
Submit
Filters
Show top: 100
Hide common words
Zoom
In Out Reset
Showing 73 of 1719 terms
flushed
coughing
laughter
lips
mother
repeated
eyes
opened
made
voice
soul
heart
face
body
head
neck
blood
door
gave
shame
rage
father
children
heard
hair
power
love
mr
smiled
water
fleming
stephen
charles
turned
rose
fell
sweet
darkness
black
sad
terrible
blue
white
soft
air
beautiful
long
silence
cried
gentle
simple
strange
cruel
gloom
high
low
pale
cold
raised
thought
red
holy
damp
unfair
green
dark
holly
warm
silent
young
small
weak
humble
Portrait of the Artist as a Young Man
X and Y

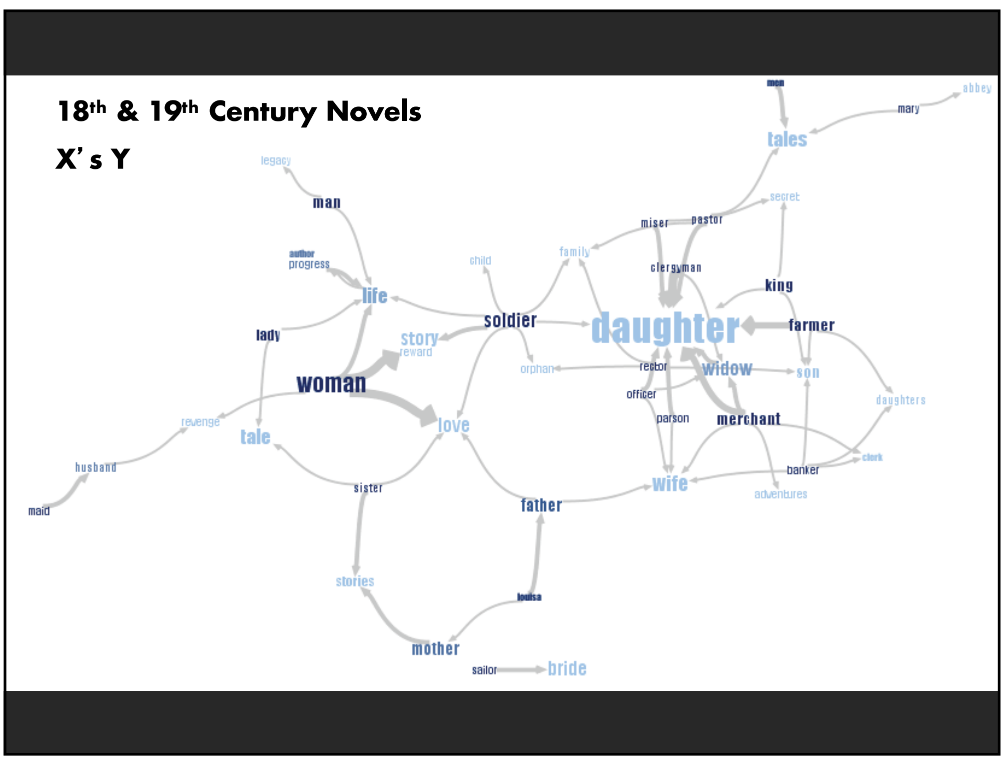
72


The Bible
X begat Y
shaharaim
pharez
children
hezron
ram
amminadab
abraham
nahshon
aminadab
aram
esrom
joseph
salma
salmon
naasson
phares
judas
isaac
boaz
attai
nathan
zabad
ephlal
booz
david
jacob
obed
jesse
matthan
johanan
jehu
eleazar
eliud
helez
azariah
phinehas
achim
amariah
abishua
sadoc
eleasah
hilkiah
bukki
azor
seraiah
meraioth
uzzi
eliakim
abiud
zorobabel
sisamai
zerahiah
ahimaaz
ahitub
shallum
jekamiah
zadok
elishama
joab
jehozadak

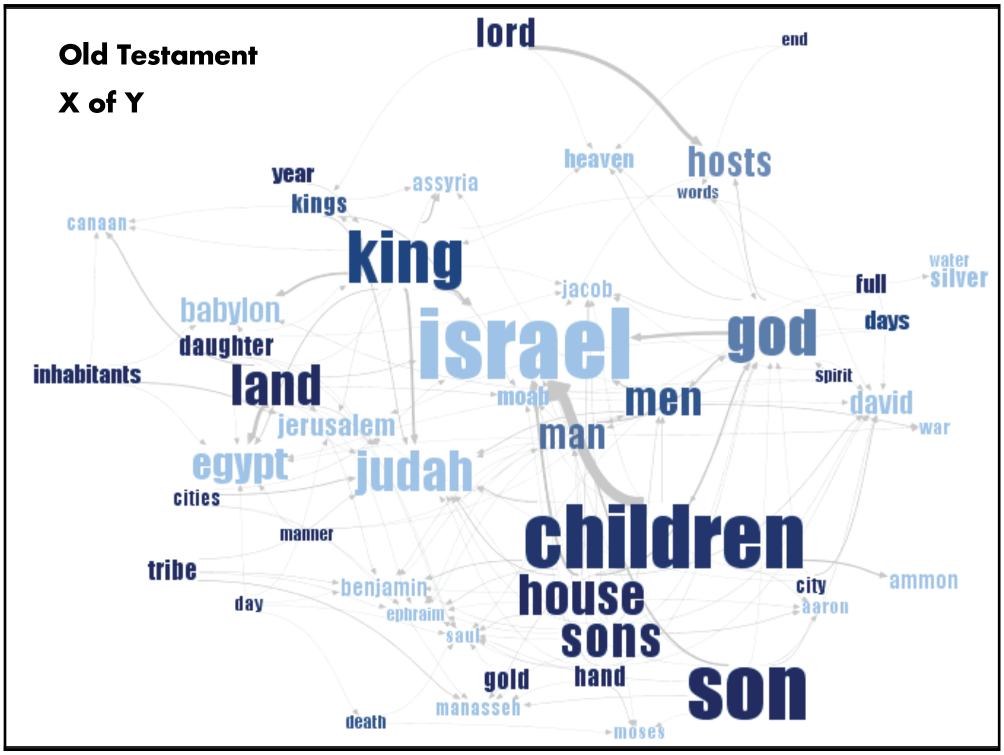
74


Pride & Prejudice
X at Y
darcy
glance
jane
looked
attempt
bennet
peep
elizabeth
present
conversation
party
arrival
staying
uncle
pemberley
home
lambton
passed
appearance
christmas
reception
whist
play
stay
presentation
st
hunsford
friends
settled
remained
week
longbourn
dance
relations
dine
ladies
court
netherfield
rosings
meryton
ball
remaining
dining
engagements
room
table
night
evening
mr

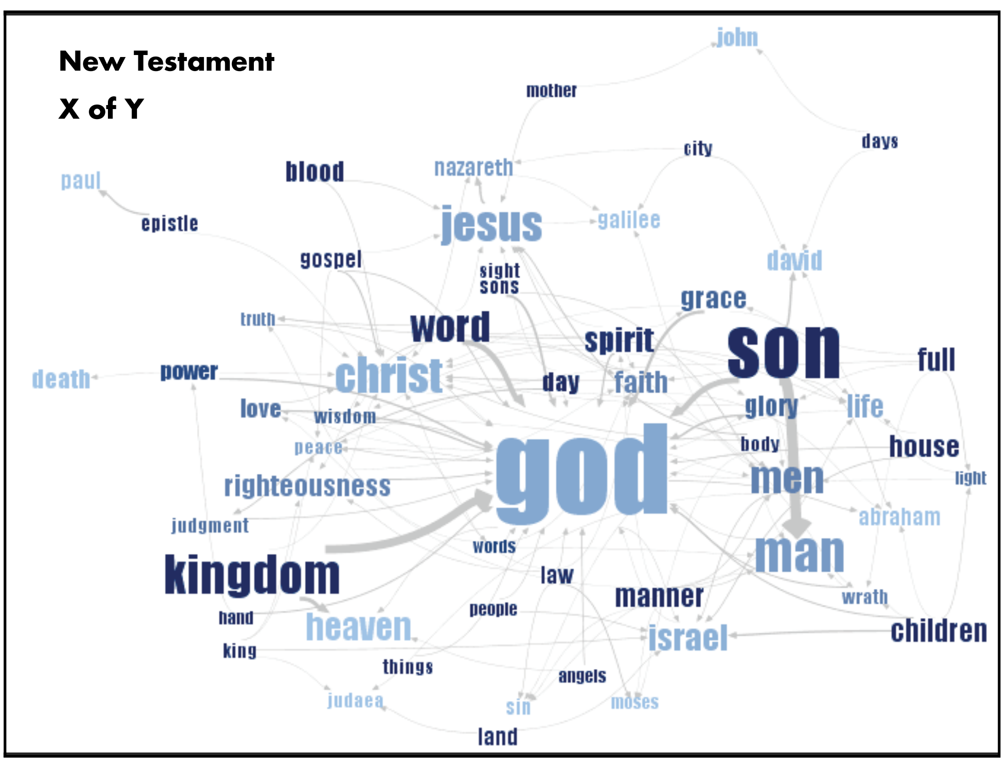
75


18th & 19th Century Novels
X's Y
men
abbey
mary
tales
legacy
man
secret
miser
pastor
author
progress
child
family
clergyman
king
life
soldier
daughter
farmer
lady
story
reward
orphan
rector
widow
son
woman
officer
daughters
revenge
love
parson
merchant
tale
clerk
husband
banker
sister
wife
adventures
maid
father
stories
louisa
mother
sailor
bride

78


Old Testament
X of Y
lord
end
heaven
hosts
words
year
kings
assyria
canaan
king
jacob
water
silver
full
days
babylon
daughter
inhabitants
land
israel
god
spirit
moab
men
david
war
jerusalem
man
egypt
judah
cities
manner
children
tribe
benjamin
city
ammon
house
aaron
day
ephraim
saul
sons
gold
hand
son
manasseh
death
moses

79


New Testament
X of Y
john
mother
days
city
paul
blood
nazareth
epistle
jesus
galilee
gospel
david
sight
sons
grace
truth
word
spirit
son
full
death
power
christ
day
faith
love
glory
life
wisdom
body
house
peace
god
men
light
righteousness
abraham
judgment
man
words
kingdom
law
wrath
manner
hand
heaven
people
israel
children
king
things
angels
judaea
sin
moses
land

80

Visualizing Conversation

91

Visualizing Conversation

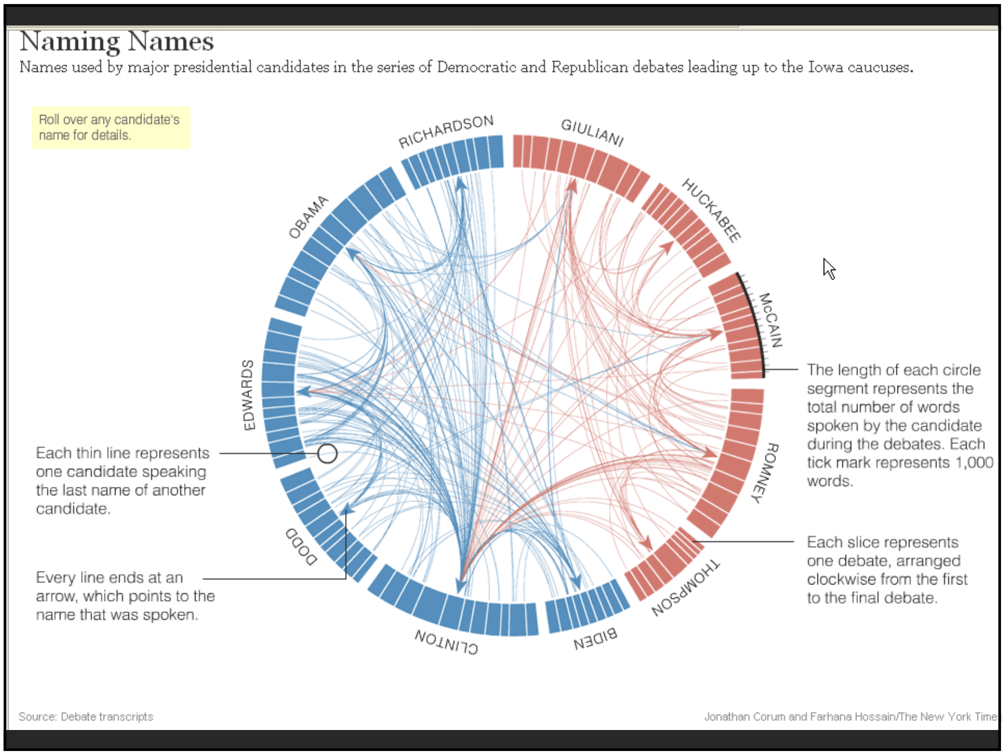
Many dimensions to consider:

- 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

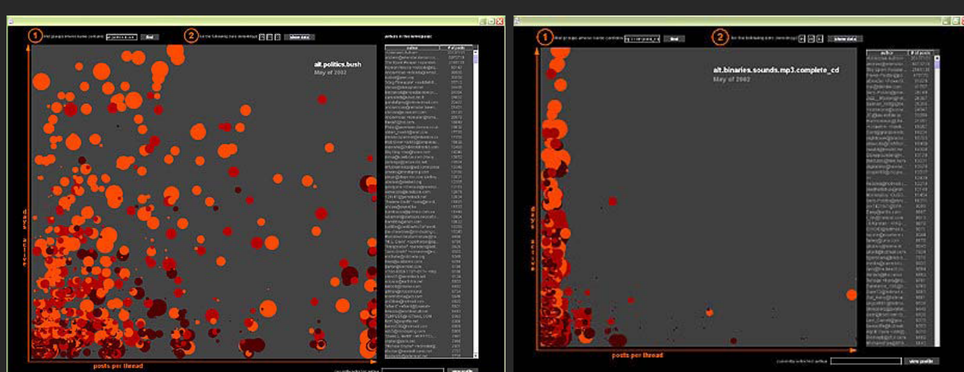
92



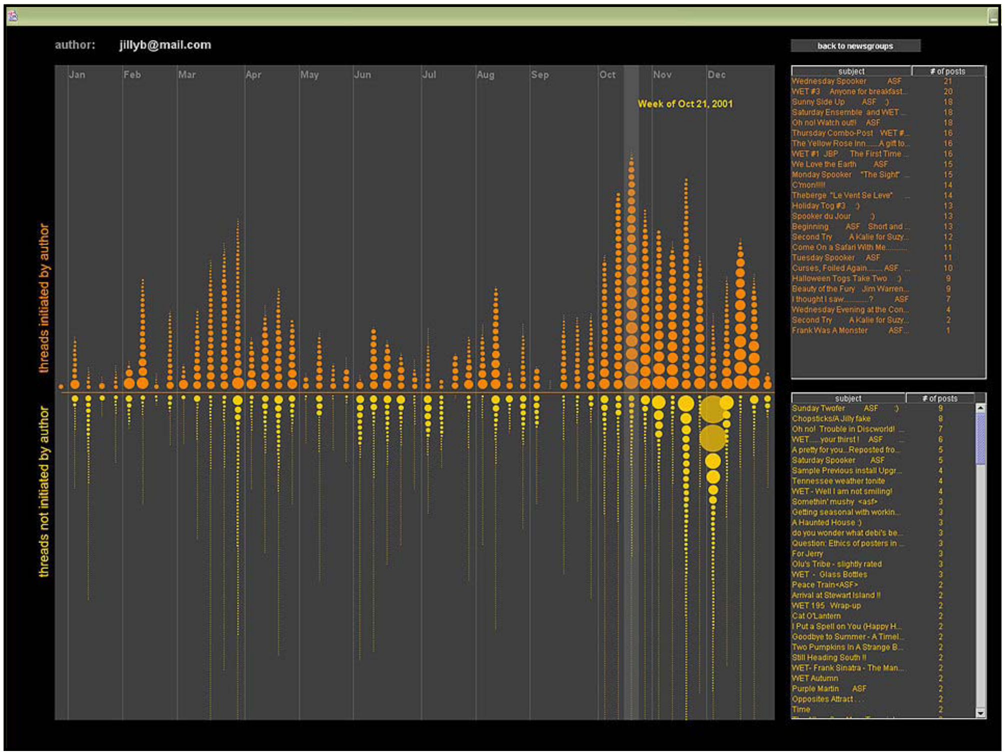
93

Usenet Visualization [Viégas]

Show correspondence patterns in text forums
Initiate vs. reply; size and duration of discussion



96


author: jillyb@mail.com
back to newsgroups
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Week of Oct 21, 2001
threads initiated by author
threads not initiated by author
subject
# of posts

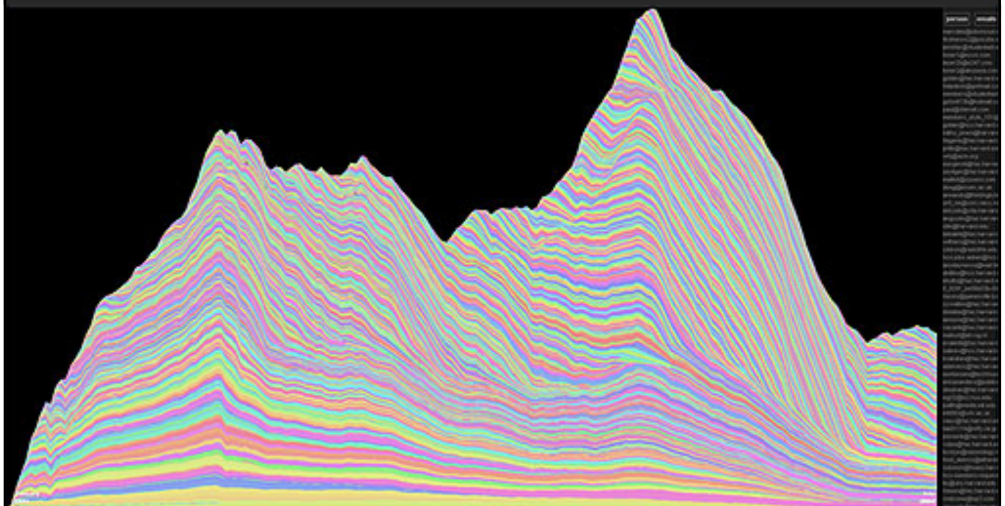
97


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?
threads initiated by author
threads not initiated by author
subject
# of posts

98

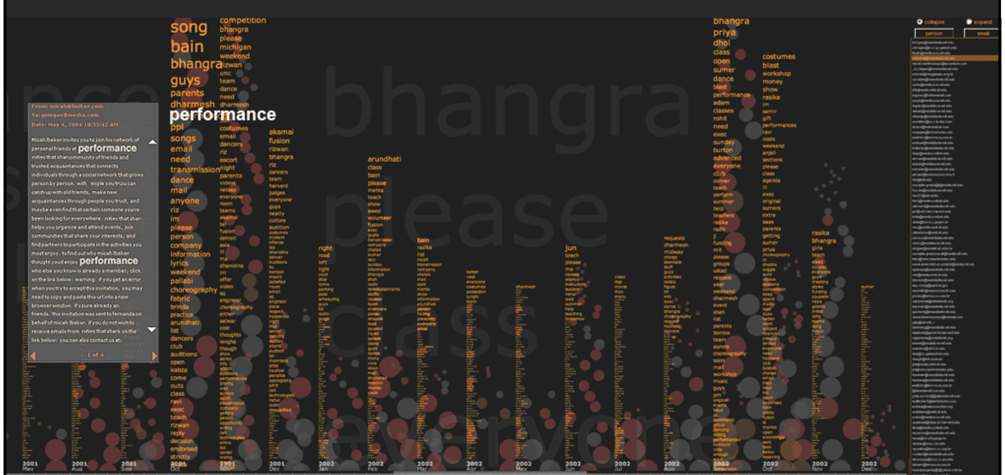
Mountain (Viégas)

Conversation by person over time (who x when)



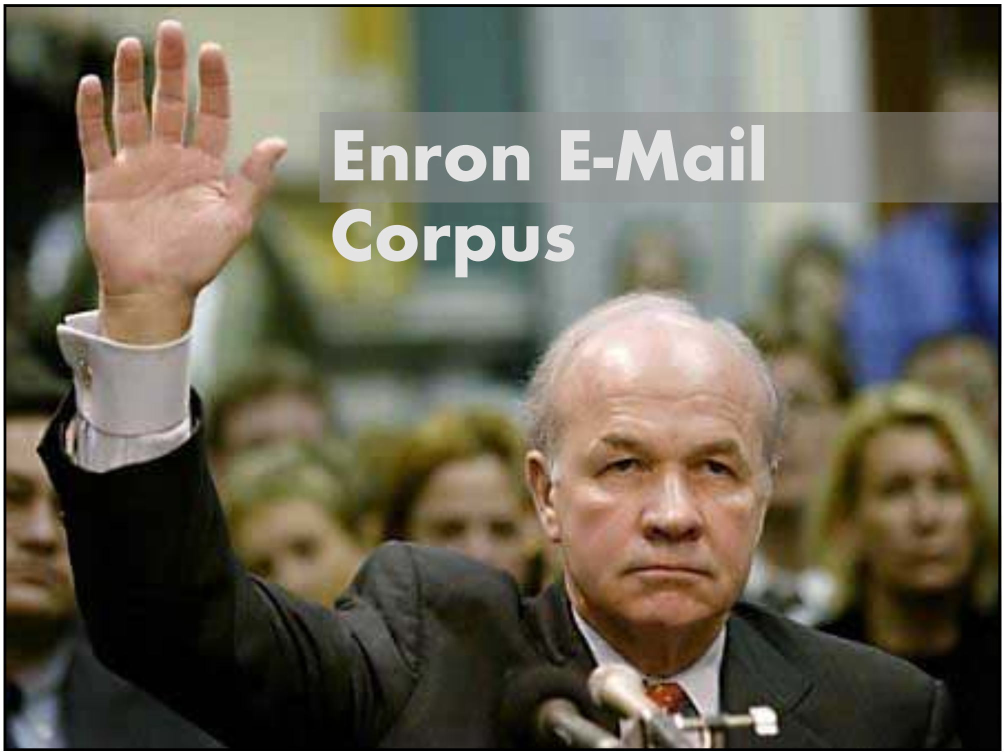
99

Themail (Viégas)

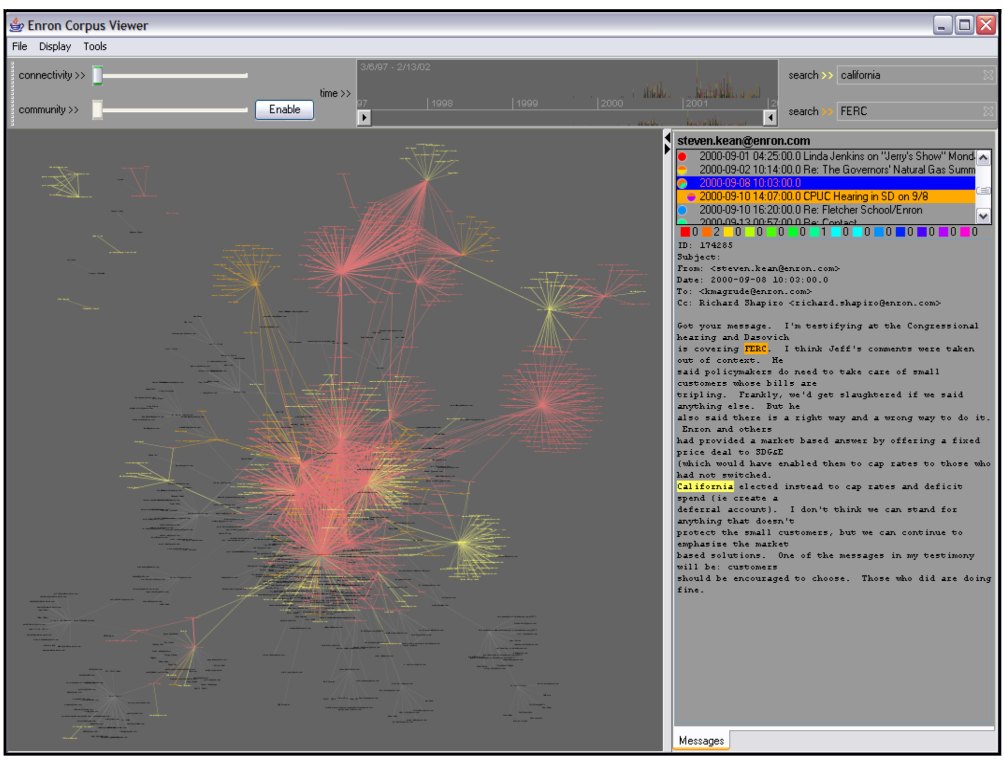


One person over time, TF.IDF weighted terms

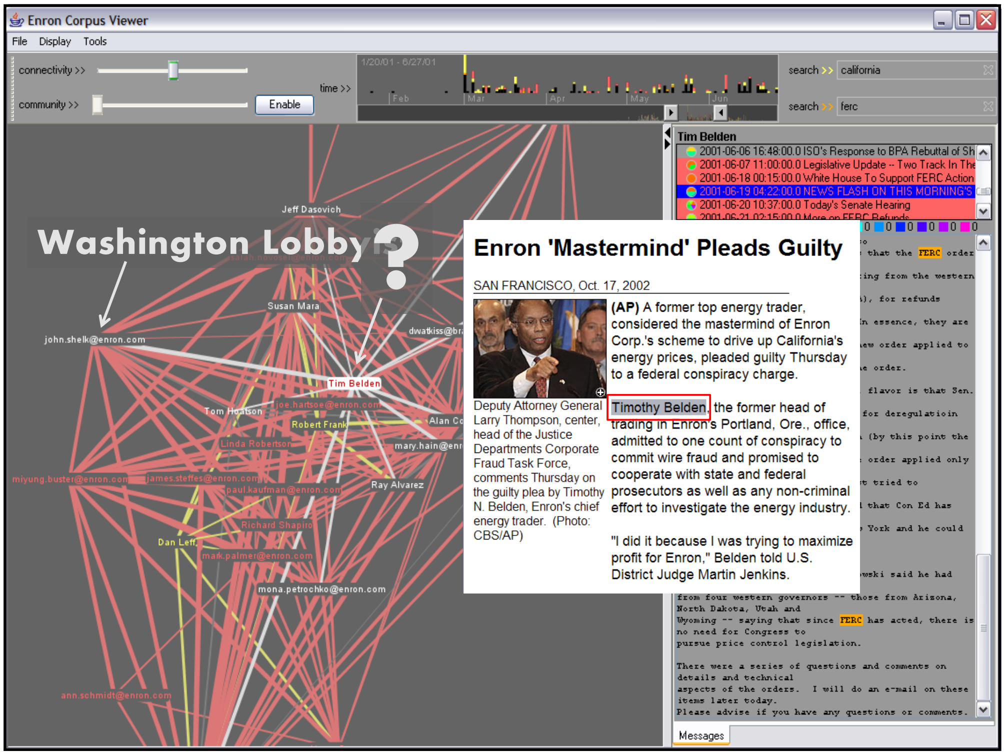
100


Enron E-Mail Corpus

101


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 1 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

102



103

Visualizing Document Collections

104

Named Entity Recognition

Identify and classify named entities in text:

John Smith → PERSON

Soviet Union → COUNTRY

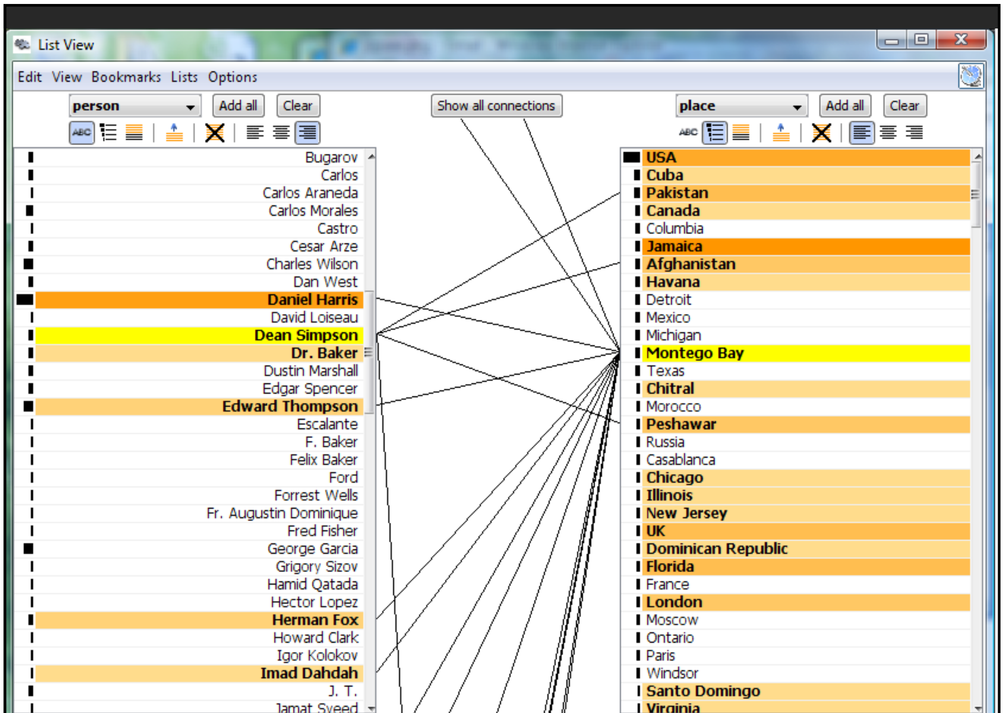
353 Serra St → ADDRESS

(555) 721-4312 → PHONE NUMBER

Entity relations: how do the entities relate?

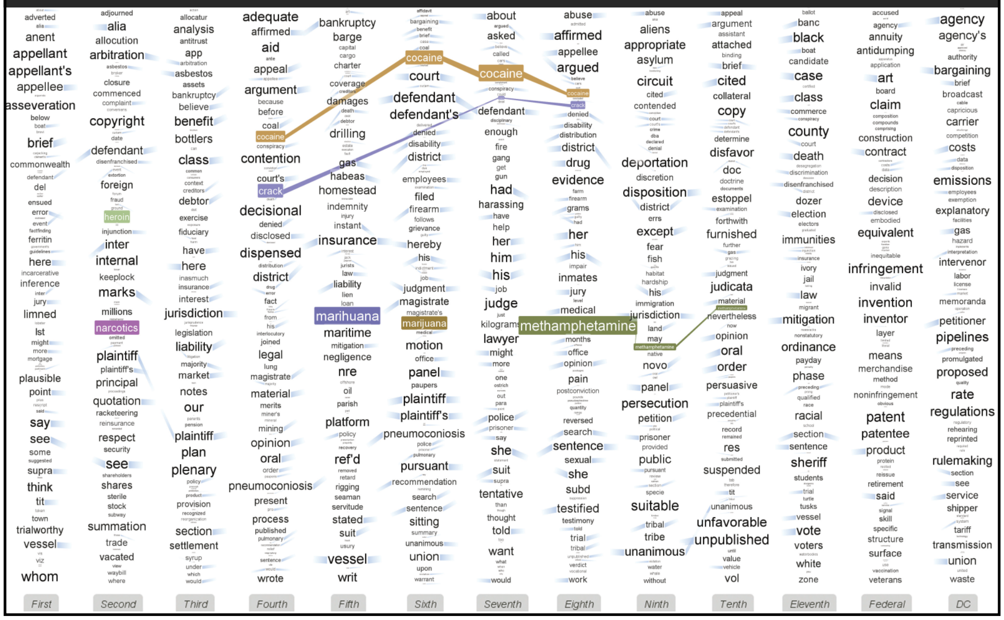
Simple approach: do they co-occur in small window of text?

107



108

Parallel Tag Clouds [Collins 09]



109

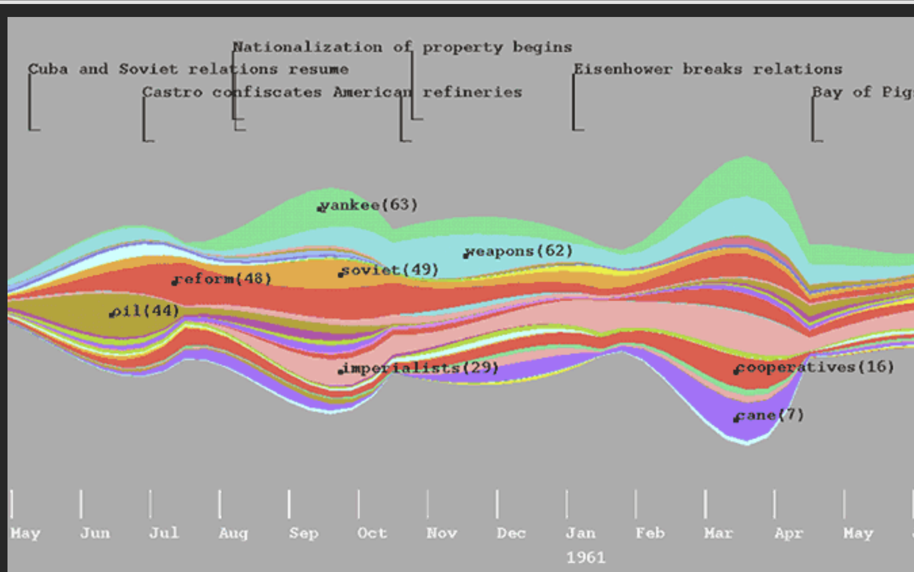
Topic modeling

Topic modeling approaches

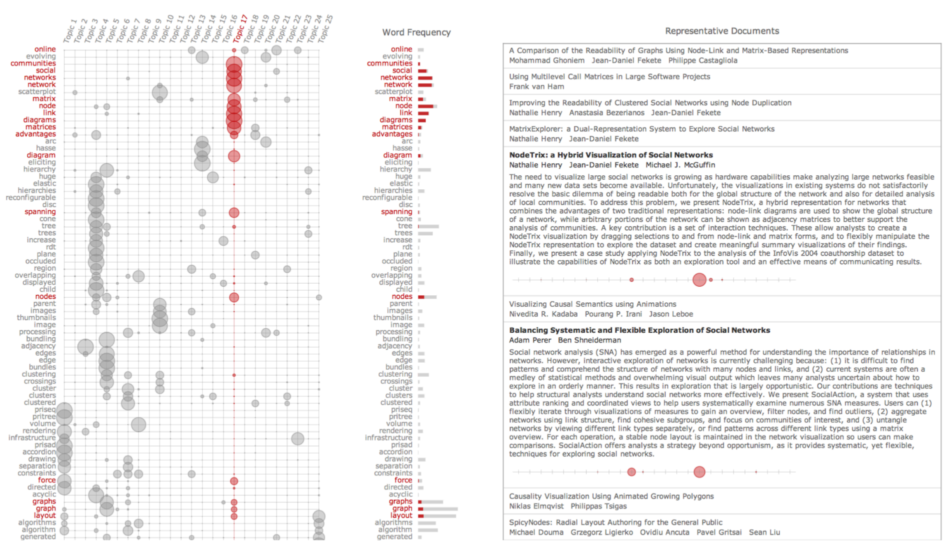
- Assume documents are a mixture of topics
- Topics are (roughly) a set of co-occurring terms
- Latent Semantic Analysis (LSA): reduce term matrix
- Latent Dirichlet Allocation (LDA): statistical model

113

ThemeRiver (Havre et al 99)



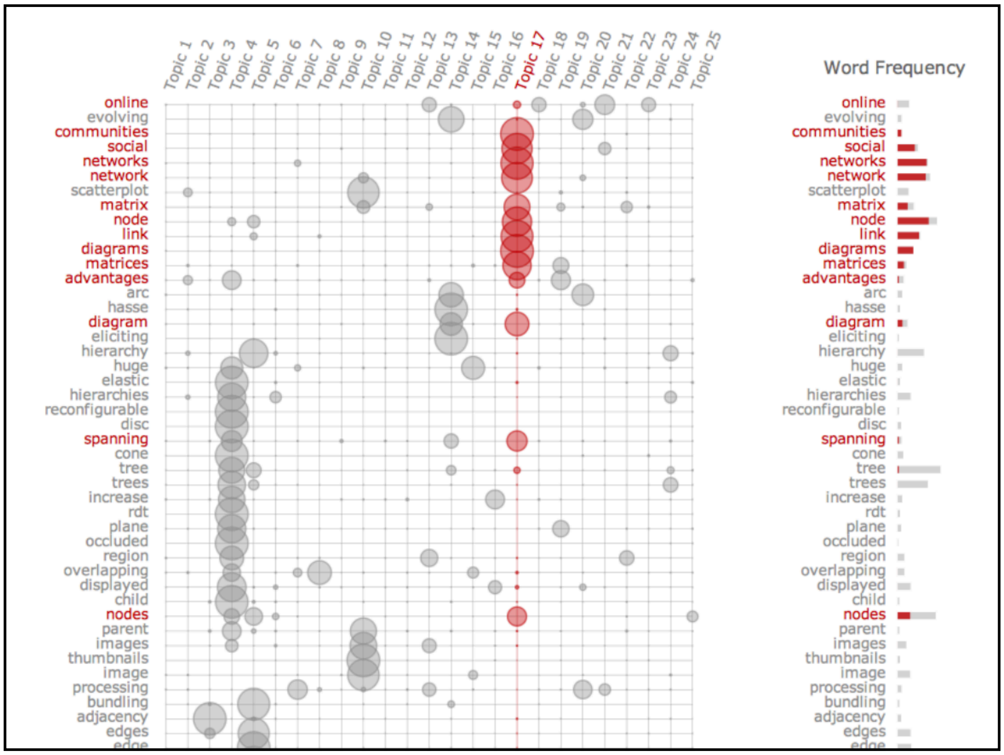
114



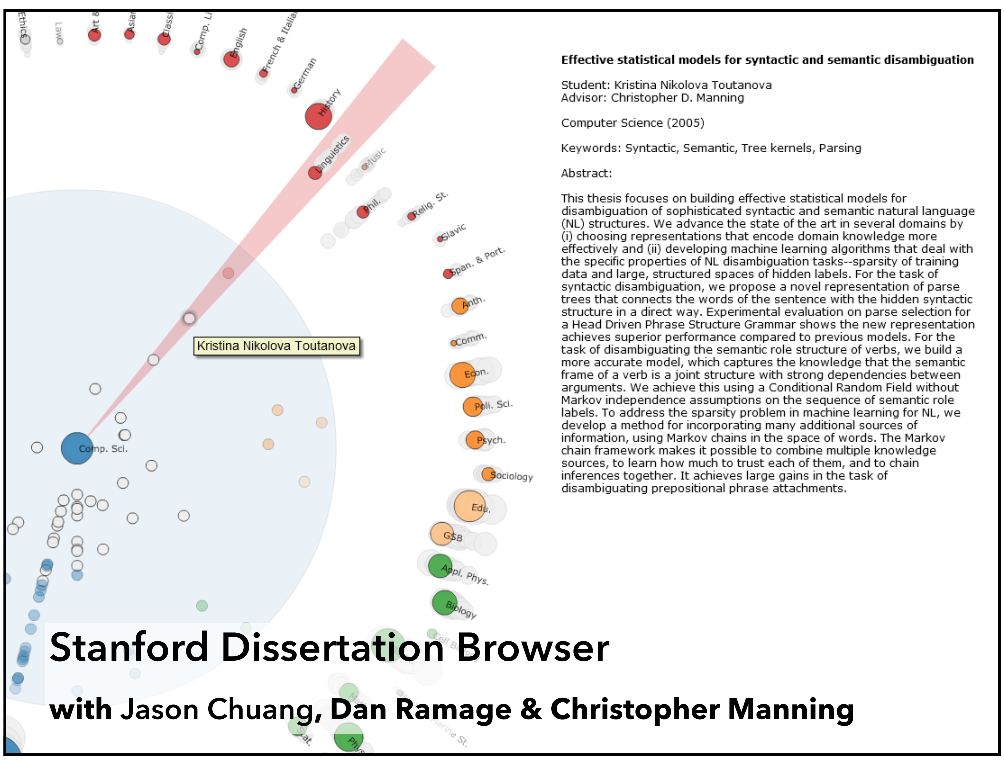
Termite: Visualizing Topic Models [Chuang '12]

Show salient (vs. frequent) terms. Seriate rows & columns.

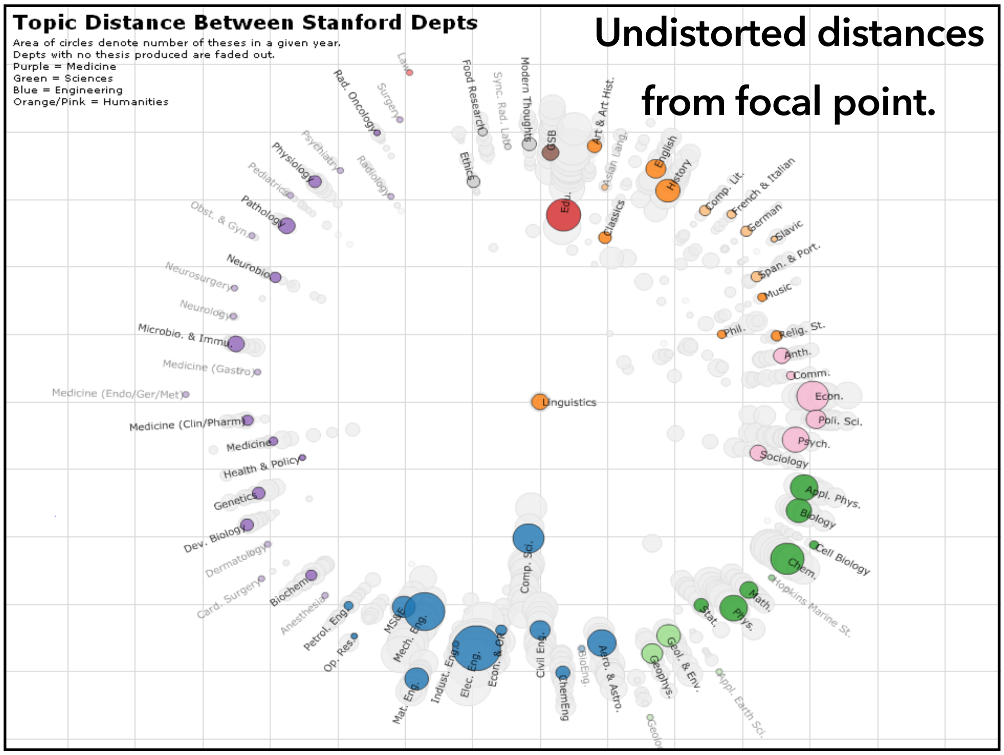
122


Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8
Topic 9
Topic 10
Topic 11
Topic 12
Topic 13
Topic 14
Topic 15
Topic 16
Topic 17
Topic 18
Topic 19
Topic 20
Topic 21
Topic 22
Topic 23
Topic 24
Topic 25
Word Frequency
online
evolving
communities
social
networks
network
scatterplot
matrix
node
link
diagrams
matrices
advantages
arc
hasse
diagram
eliciting
hierarchy
huge
elastic
hierarchies
reconfigurable
disc
spanning
cone
tree
trees
increase
rdt
plane
occluded
region
overlapping
displayed
child
nodes
parent
images
thumbnails
image
processing
bundling
adjacency
edges

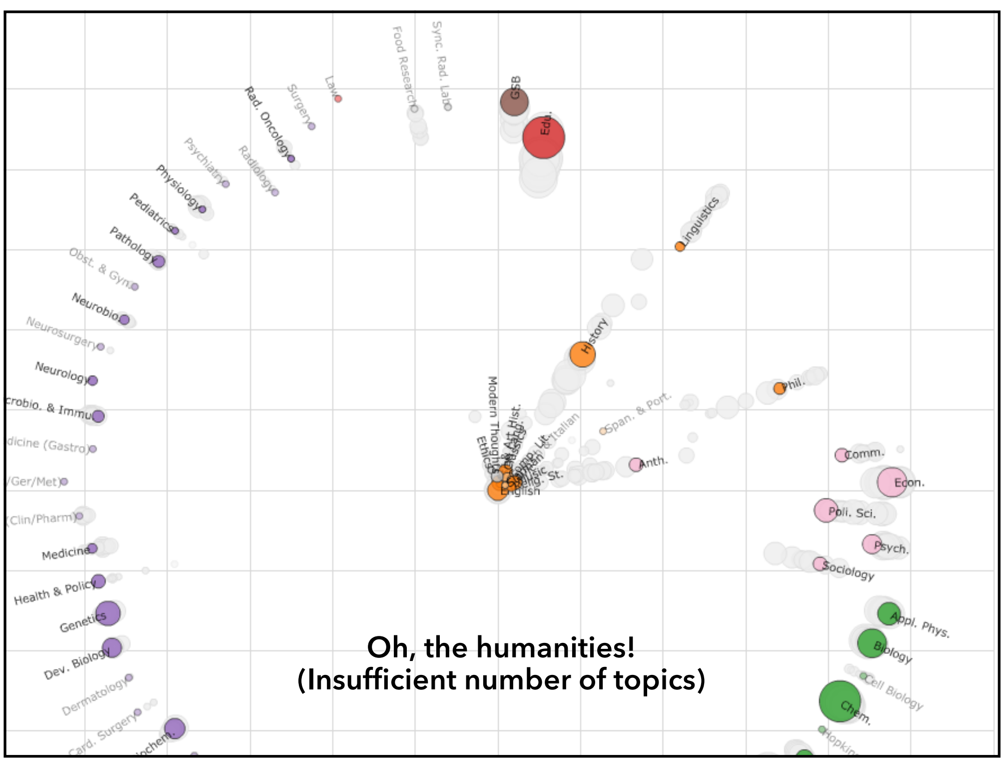
123


English
French & Italian
German
History
Linguistics
Music
Phil.
Relig. St.
Slavic
Span. & Port.
Anth.
Comm.
Econ.
Poli. Sci.
Psych.
Sociology
Edu.
GSB
Appl. Phys.
Biology
Comp. Sci.
Kristina Nikolova Toutanova
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

124


Topic Distance Between Stanford Depts
Area of circles denote number of theses in a given year.
Depts with no thesis produced are faded out.
Purple = Medicine
Green = Sciences
Blue = Engineering
Orange/Pink = Humanities
Undistorted distances
from focal point.
Linguistics
Econ.
Poli. Sci.
Psych.
Sociology
Comp. Sci.
English
History
Classics
GSB
Ethics
Food Research
Modern Thought
Art & Art Hist.
Microbio. & Immu.
Medicine (Clin/Pharm)
Medicine
Health & Policy
Genetics
Dev. Biology
Biochem.
Appl. Phys.
Biology
Cell Biology
Chem.
Math
Phys.
Stat.
Geophys.
Civil Eng.
ChemEng
Mat. Eng.
Indust. Eng.
Elec. Eng.
Mech. Eng.
Petrol. Eng.
Op. Res.

126


Oh, the humanities!
(Insufficient number of topics)
GSB
Edu.
Linguistics
History
Phil.
Anth.
Comm.
Econ.
Poli. Sci.
Psych.
Sociology
Appl. Phys.
Biology
Cell Biology
Chem.
Rad. Oncology
Physiology
Pediatrics
Pathology
Neurobio.
Neurology
Medicine
Health & Policy
Genetics
Dev. Biology

127

Summary

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

Understand abstraction of your language models.
Match analysis task with appropriate tools & models.

Currently: from bag-of-words to *vector space embeddings*

134