CSE 512 - Data Visualization Text Visualization



Jeffrey Heer University of Washington

Why Visualize Text?

Why Visualize Text?

Understanding – get the "gist" of a document

Grouping – cluster for overview or classification

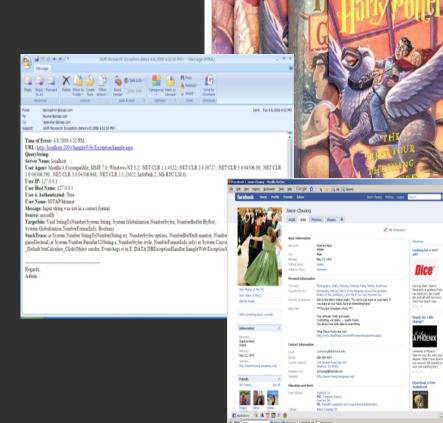
Comparison – compare document collections, or inspect evolution of collection over time

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

Text as 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: Health Care Reform

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.

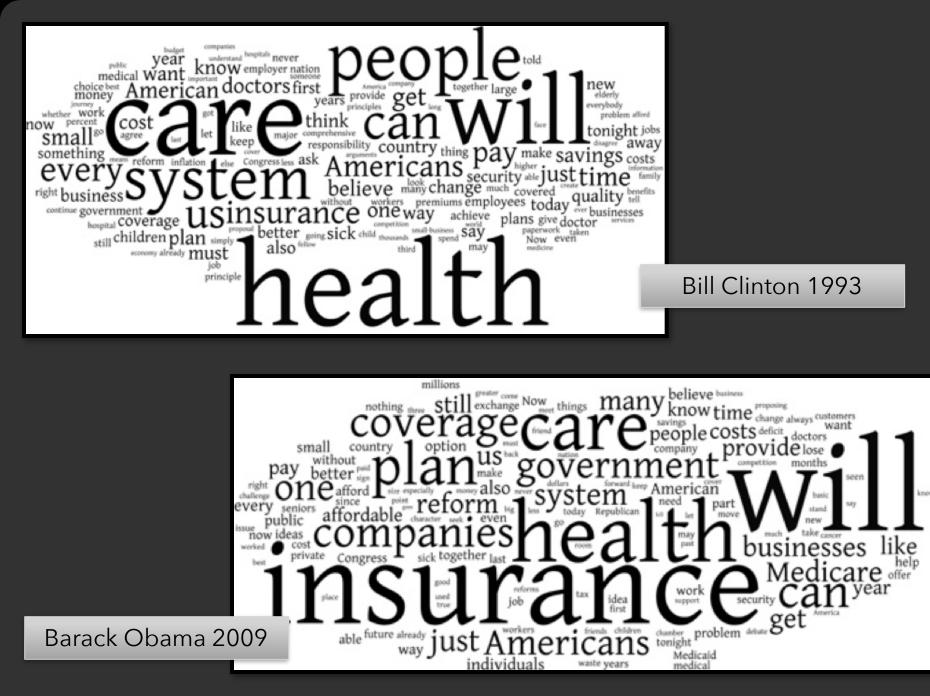
But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of yo

Tag Clouds: Word Count

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



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



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

Word Tree: Word Sequences

Visualizations : Word Tree President Obama's Address to Congress on Health Care

Search 🚺	Back Forward Start End Occurrence Order Clicks Will Zoom
52 hits	<pre>spectrum control in the spectrum control integration of the spectrum control integration cont</pre>
	and the failer than making and charms about a government biak and the strateging and the set in the strateging and the set in the

Search i will	Back Forward 💿 Start 🖸 End Occurrence Order 💠 Clicks Will Zoom
12 hits	let up until those americans who seek jobs can find them (applause) until tho
	back down on the basic principle that if americans can't find affordable coverage , w
	a plan that adds one dime to our deficits either now or in the future
	notit if it adds one dime to the deficit , now or in the future , period
i savill /	waste time with those who have made the calculation that it's better politics to kill thi
i will	and i will not accept the status quo as a solution .
	accept the status quo as a solution .
	make sure that no government bureaucrat or insurance company bureaucrat gets between you and t
	protect medicare .
	continue to seek common ground in the weeks ahead .
	be there to listen .

still believe

We can extreme the hard . - i still believe that we can act when it's hard . - i still believe that we can act when it's hard . - that we can act when it's hard .

Gulfs of Evaluation

Many 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?

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.

Topics

Text as Data

Visualizing Document Content Evolving Documents Visualizing Conversation Document Collections

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 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

Tokenization & 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	

WordCounts (Harris '04)

			WORDCOUNT
PREVIOUS WORD			NEXT WORD
the	ofandtoainthatitiswasi.for	ບທັດດາກິດຕົດທີ່ຖຸມຕິດອີກຈີເມືອກອີດເຊີຍເຊີຍເຊືອດອີກເຊືອດອີກເຊືອດອີກເຊືອດອີກເຊືອດອີກເຊືອດອີກເຊືອດອີກເຊືອດອີກເຊືອດ	1977 (Ball) (Bal
CURRENT WORD			
FIND WORD:	BY RANK:	JESTED WORD: THE	86800 WORDS IN ARCHIVE
		RANK: 1	ABOUT WORDCOUNT

http://wordcount.org

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

Creator: Anonymous Tags:

Edit Language Font Layout Color



Tag 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

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 $tf.idf_{td} = log(1 + tf_{td}) \times log(N/df_t)$ $df_t = # docs containing t; N = # of docs$

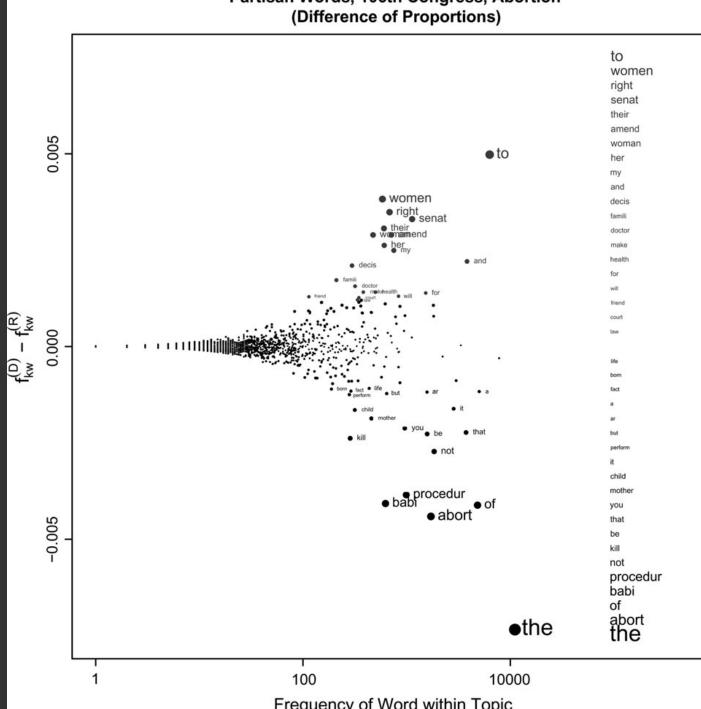
Keyword Weighting

Term Frequency

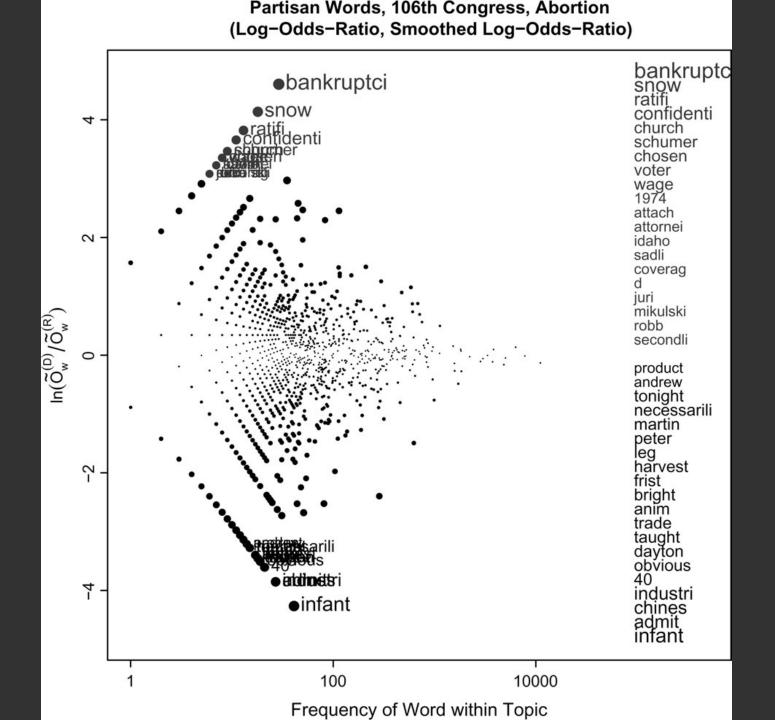
 $tf_{td} = count(t) in d$

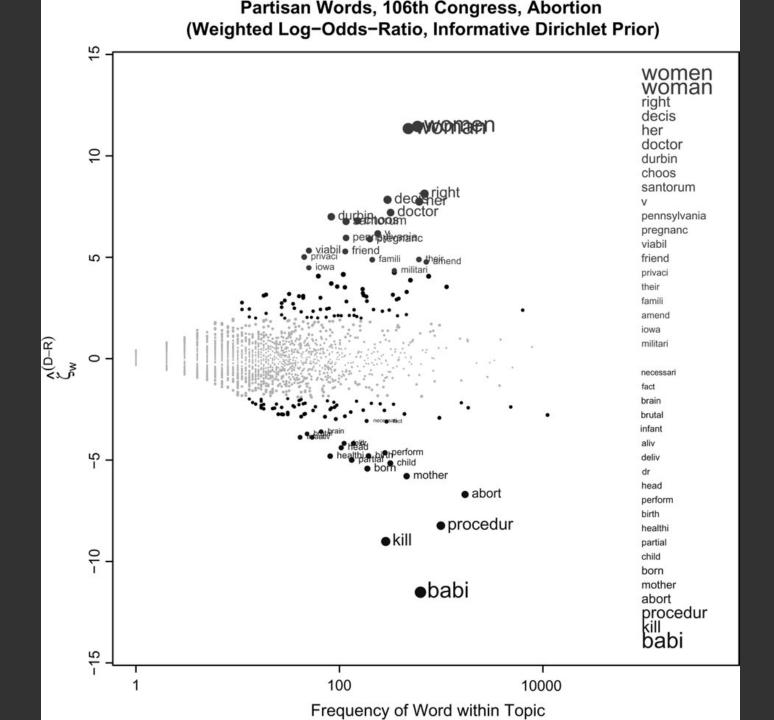
TF.IDF: Term Freq by Inverse Document Freq $tf.idf_{td} = log(1 + tf_{td}) \times log(N/df_{t})$ $df_{t} = # docs containing t; N = # of docs$

 $\begin{aligned} \mathbf{G}^{2}: \mathbf{Probability of different word frequency} \\ \mathbf{E}_{1} &= |\mathbf{d}| \times (\mathrm{tf}_{\mathrm{td}} + \mathrm{tf}_{\mathrm{t(C-d)}}) / |\mathbf{C}| \\ \mathbf{E}_{2} &= |\mathbf{C} \cdot \mathbf{d}| \times (\mathrm{tf}_{\mathrm{td}} + \mathrm{tf}_{\mathrm{t(C-d)}}) / |\mathbf{C}| \\ \mathbf{G}^{2} &= 2 \times (\mathrm{tf}_{\mathrm{td}} \log(\mathrm{tf}_{\mathrm{td}}/\mathrm{E}_{1}) + \mathrm{tf}_{\mathrm{t(C-d)}}\log(\mathrm{tf}_{\mathrm{t(C-d)}}/\mathrm{E}_{2})) \end{aligned}$



Partisan Words, 106th Congress, Abortion





Limitations of Freq. 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

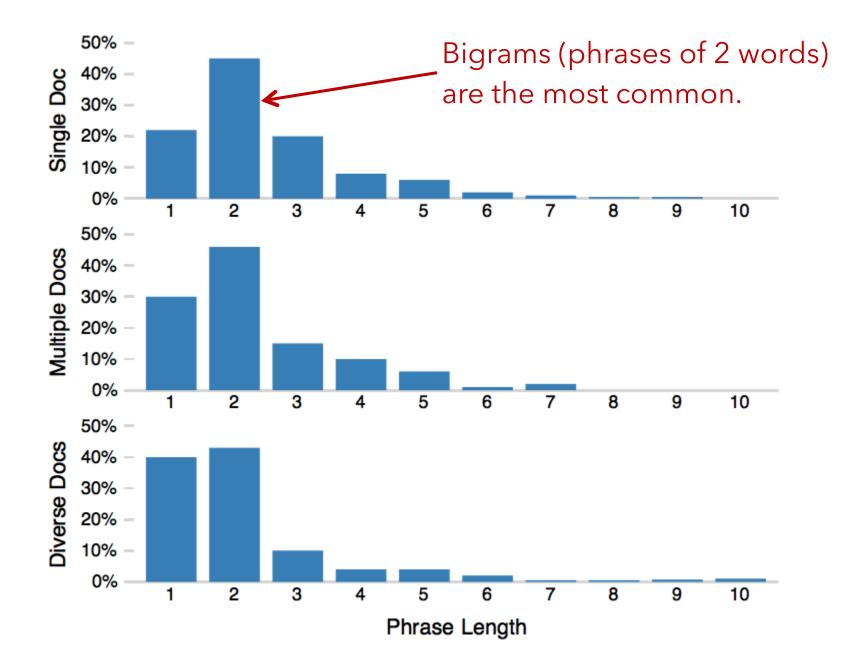
How do people describe text?

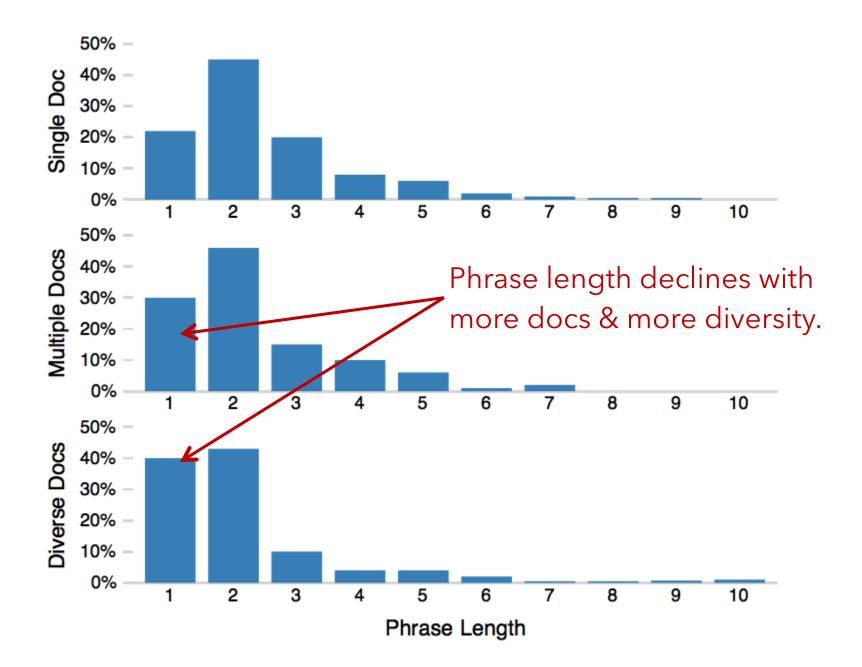
We asked 69 subjects (graduate students) to read and describe dissertation abstracts.

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, Manning & Heer, 2012]



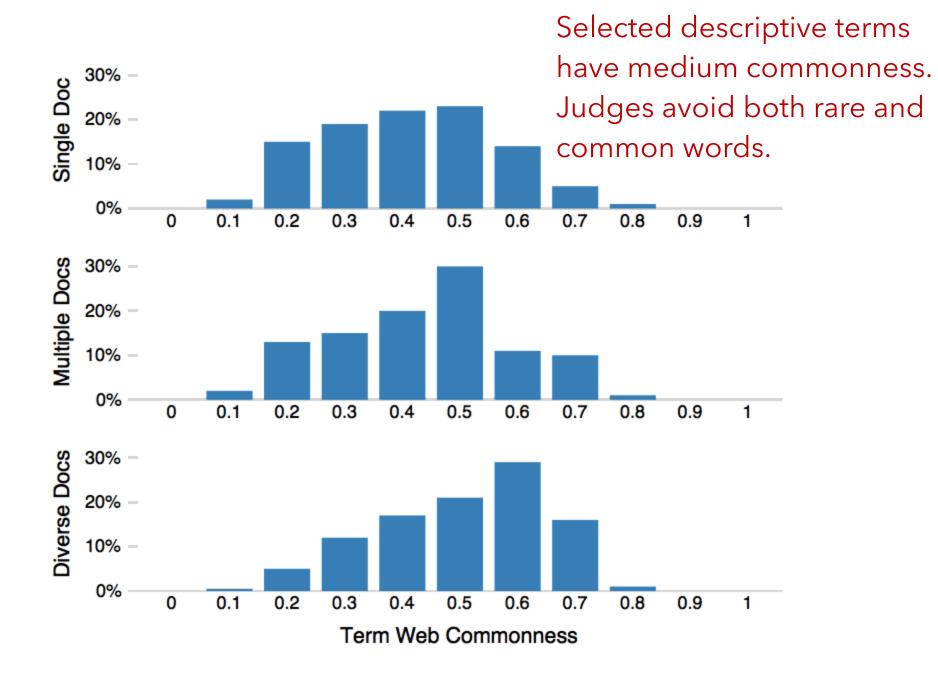


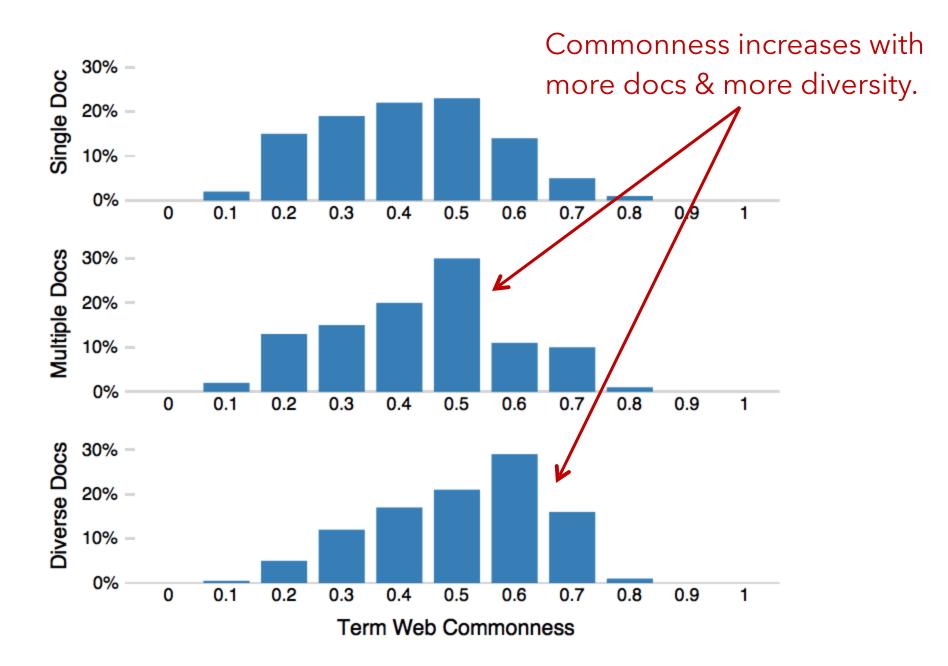
Term Commonness

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

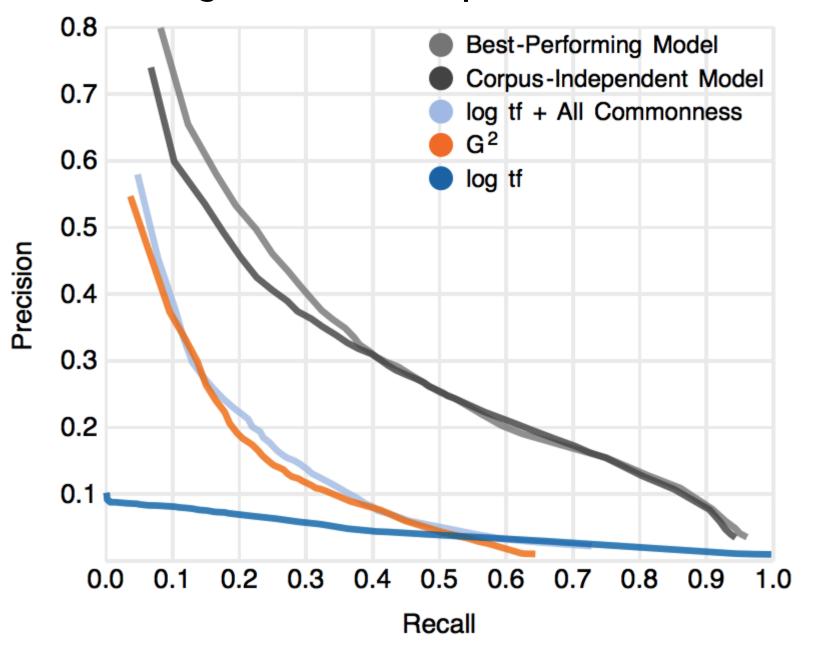
The normalized term frequency relative to the most frequent n-gram, e.g., the word "the".

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





Scoring Terms with Freq, Grammar & Position



🔆 🕂 🖗 http://www.boeing.com/Features/2010/09/bds_feat_morewater 🖒

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



CHAMILA JAYAWEERA/BOEING

Q-

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.

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.

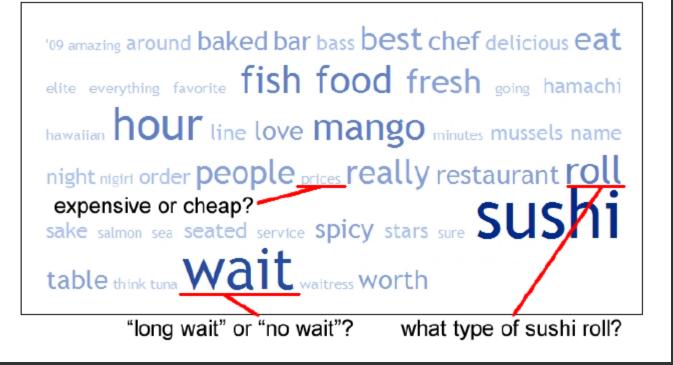


fighter F/A Hornet Super Boeing -18 rain St. jet Louis 15-inch-per-hour douse hangar water-tight Check Baxter sea-based aircraft Rich seep click Navy sure Water moisture watch

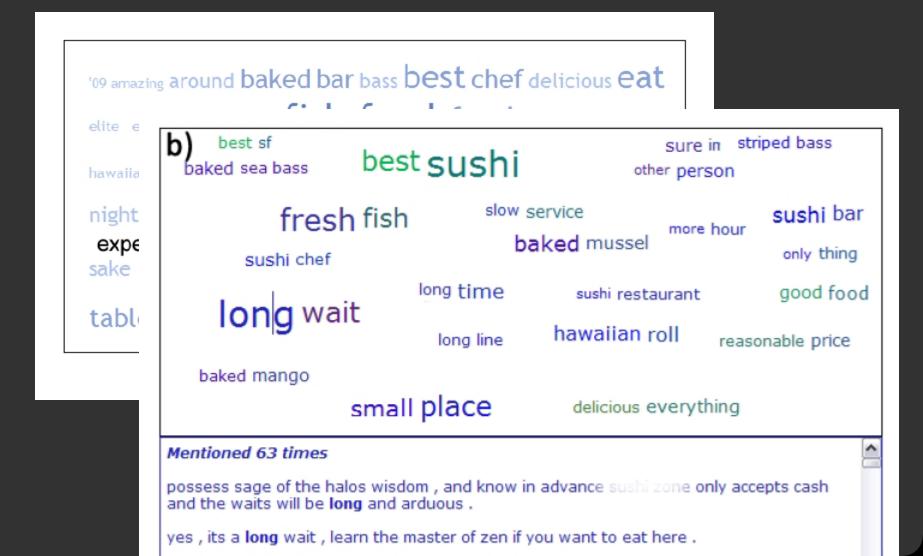
enormous stay want

Regression Model Super Hornet F/A -18 fighter jet **Boeing engineers** special process rain check electronics suite Program experts simulated rain ultimate customers enormous hangar water-tight integrity Rich Baxter 15-inch-per-hour rate video story aircraft U.S. Navy fighter pilots Super Hornet final assembly manager U.S. Navyfighter tighter pilot sea-based tighter

Yelp Review Spotlight (Yatani 2011)



Yelp Review Spotlight (Yatani 2011)



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

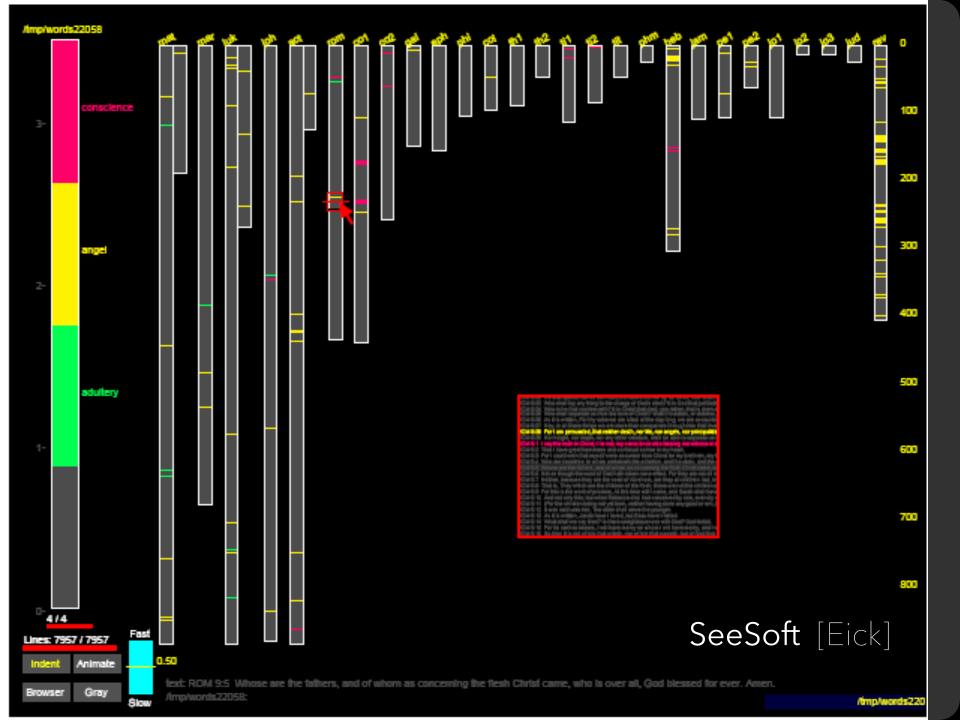
Information Retrieval

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

Google scholar acronym resolution Search Advanced Scholar Search	
Scholar Articles and patents 💌 anytime 💌 include citations 💌 🔀 Create email alert	
A supervised learning approach to acronym identification	[PDF] from nrc-cnrc.
D Nadeau, P Turney - The Eighteenth Canadian, 2005 - nparc.cisti-icist.nrc-cnrc.gc.ca Recently the fields of Genetics and Medicine have become especially interested in acronym resolution (Pustejovsky et al., 2001, Yu et al. 2002) Pustejovsky et al.'s acronym resolution technique searches for definitions of acronyms within noun phrases <u>Cited by 48 - Related articles - All 16 versions</u>	
Biomedical term mapping databases JD Wren, JT Chang, J Pustejovsky Nucleic acids, 2005 - Oxford Univ Press the prevalence of polynyms, or acronyms with multiple definitions. An important part of any high-throughput effort to tie experimental findings to published knowledge within the scientific literature involves acronym resolution Cited by 41 - Related articles - All 22 versions	[HTML] from nih.gov Find it@Stanford
Anthropogenic climate change over the Mediterranean region simulated by a global variable resolution	Find it@Stanford
model AL Gibelin Climate Dynamics, 2003 - Springer	
The long simulations CC and CS are split into two 30-year datasets CC1 and CS1 for the period 1960–1989 and CC2 and CS2 for the period 2070–2099 Full name Acronym	
Resolution Period Coupled Coupled control CC T63 1950-2099 Yes	
<u>Cited by 197</u> - <u>Related articles</u> - <u>BL Direct</u> - <u>All 5 versions</u>	
Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare	
enterprises.	
MS Tuttle, NE Olson, KD Keck, WG Cole Methods of information, 1998 - ukpmc.ac.uk Title not supplied (PMID:10566483). Concept definition and manipulation are supported through	

User Query (Enter words for different topics or	n different lines.)	Run Search New Query Quit		
osteoporosis		Search Limit: 📀 50 📀 100 🔶 250 📀 500 × 100		
prevention research		Number of Clusters: \bigcirc 3 \bigcirc 4 \blacklozenge 5 \bigcirc 8 \bigcirc 10		
Mode: TileBars				
Cluster Titles		Backup		
	FR88513-0	0157		
	AP: Groups Seek \$1 Billion a Year for Aging Research			
	SJMN: WOMEN'S HEALTH LEGISLATION PROPOSED CF			
	AP: Older Athletes Run For Science			
Barris Barrison and Salah	FR: Commi	ittee Meetings		
Building a descent	FR: October Advisory Committees; Meetings			
I a sharehold a farmer has been	FR88120-0046			
	FR: Chronic Disease Burden and Prevention Models; Program 2			
	AP: Survey Says Experts Split on Diversion of Funds for AIDS			
	FR: Consolidated Delegations of Authority for Policy Developm			
	SJMN: RES	EARCH FOR BREAST CANCER IS STUCK IN P		
	M			

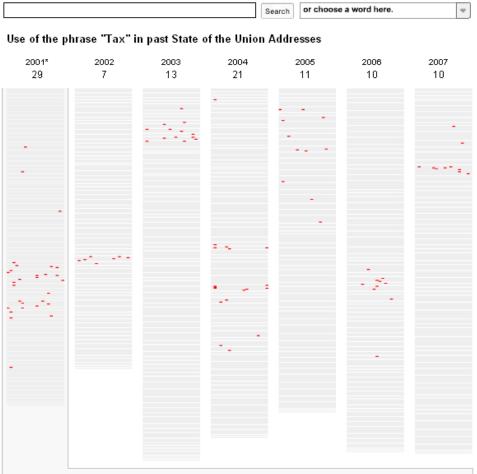
TileBars [Hearst]



THE WORDS THAT WERE USED

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.



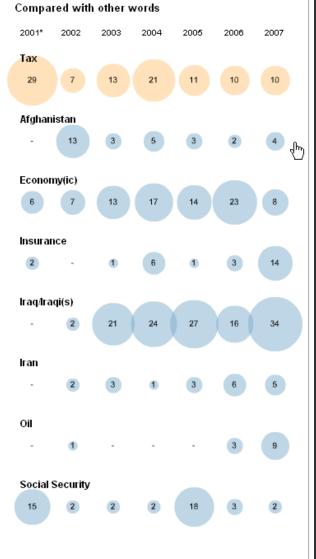
The word in context

Next Instance of 'Tax'

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.

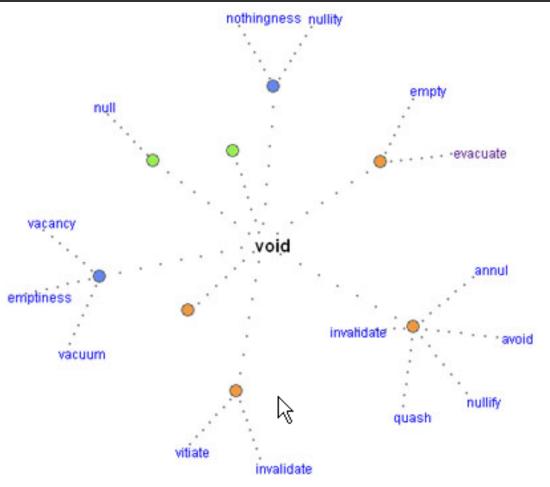
-- 2001 (Paragraph 14 of 73)

New York Times

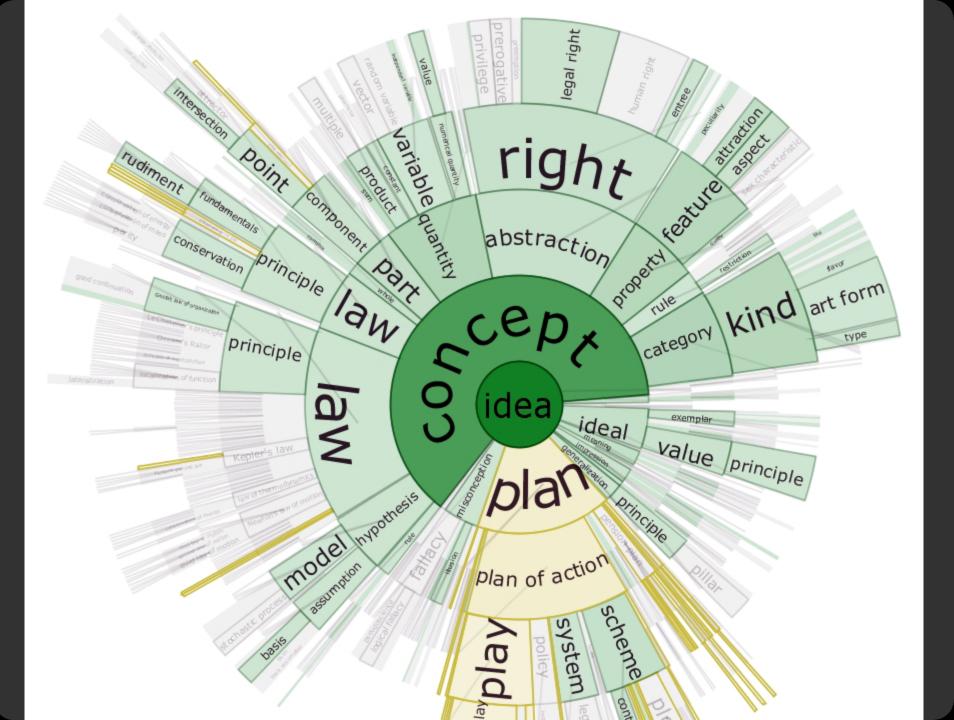


* As a newly elected president, Mr. Bush did not deliver a formal State of the Union address in 2001. His Feb. 27 speech to a joint session of Congress was analogous to the State of the Union, but without the title.

Visual Thesaurus [ThinkMap]



comments



Concordance

What is the common local context of a term?

🔏 Concordance - Larkin.Concordance						
<u>File T</u> ext <u>S</u> earch	<u>File T</u> ext <u>S</u> earch <u>E</u> dit <u>H</u> eadwords Conte <u>x</u> ts <u>V</u> iew T <u>o</u> ols Hel <u>p</u>					
🗇 Ħ 🗃 🖬 é						
Headword	No. 🔺	Context	Word	Context	Reference	
HEAR	15	That my own	heart	drifts and cries, having no	Deep Analysis	ĉ
HEARD	9	By the shout of the	heart	continually at work	And the wave	Centred
HEARING	7	Nothing to adapt the skill of the	heart	to, skill	And the wave	8
HEARS	3	The tread, the beat of it, it is my own	heart		Träumerei	
HEARSE	1	Because I follow it to my own	heart		Many famous	
HEART	25	My	heart	is ticking like the sun:	lann washed u	5
HEART'S	2	The vague	heart	sharpened to a candid co	The March Pa:	Left-aligned
HEART-SHAPED	1 📥	Contract my	heart	by looking out of date.	Lines on a Yo	ig
HEARTH	1	Having no	heart	to put aside the theft	Home is so Sa	led
HEARTS	7	And the boy puking his	heart	out in the Gents	Essential Beau	
HEARTY	1	A harbour for the	heart	against distress.	Bridge for the	
HEAT	6	These I would choose my	heart	to lead	After-Dinner F	
HEAT-HAZE	1	Time in his little cinema of the	heart		Time and Space	Index
HEATH	1	This petrified	heart	has taken,	A Stone Churc	×
HEATS	1	How should they sweep the girl clean	heart	1	l see a girl dra	
HEAVE	1	Hands that the	heart	can govern	Heaviest of flo	
HEAVEN	4	For the	heart	to be loveless, and as col	Dawn	
HEAVEN-HOLDING	1	With the unguessed-at	heart	ridi⊓g	One man walk	IZ
HEAVIER-THAN	1	If hands could free you,	heart heart	1	If hands could	None
HEAVIEST	2 🗸	That overflows the			Pour away the	2
<		<			>	
Words	Tokens	At word Deleted lines Word sort		Context sort		
7318	37070	2990 1 [24] Asc alpha	a (string)	Asc occurrence of	order	11.



Down the Flabbit-Hole

lan pictures

May

afteline Byind

Down the Rabbit-Hole

Hide text Show concordance

Show thesaurus lookup Show story line

inturbeer. ärecovered king

terconsider us unfortunate repagiputting based deeply majesty cried ^{Twinkling}hugh **DESCRIPT** hurwalk imagin@aught_enbiaded ifenktier gryphon panoihinhand wór¢: 20ice) hudles.

levau'v&urptisear&are, made indee interrupted : a lightly is .course exclaimed and most a Burganationidam and add donë begare, ethough to the golde anxigHalled inns þýlfser whisteered alice Rectastone: 印刷線 Dephains ann young young

minule haiter music wash shoulder cat dormouse iduchess sitting itself couragebre编辑创作 aawi choistory h रहती है है bawedander sidered nose Screamed appeared rénaikeø! round A Shingachully

aggs agsleep Wildignantly **B**Buple lessel hottom **Gablean**tage decidedleast blow remarkminalinaduce lives Icheshire.

Show only KWIC index (Key Word In Context)

Down the Rabbit-Hole, Rabbit with pink eyes ran close by h think it so VERY much out of the way to hear the Rabbit say to but when the Rabbt actually TOOK A WATCH OUT OF ITS WAIST COATbefore seen a rabbit with either a waistcoat-pocket, or a watch to down a large rabbit hole under the hedge.

The subbt-hale went straight on like a tunnel for some way, was another long passage, and the White Rabbit was still in comer, but the Rabor was no longer to be seen: she found It was the white Rabbit returning, splendidwidtessed, with a that she was ready to ask help of any one, so, when the Rabbit sir-1 The Rabbit started violently, dropped the white kid. curprised to see that she had put on one of the Rabbit's ittle kerselt The Babbit Sends in a Little Bit

It was the White Rabbit, trotting slowly back again, and Yery soon the Rabbit noticed Alice, as she went hunting about, messages for a rabbit 1 suppose Dinah'll be sending me on by mice and rabbits. I almost wish I hadn't gone down that abbit-hole--and yet--and yet--it's rather curious, you know, stairs. Alice knew it was the Rabbit coming to lock 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; lancied 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 Rabbill say, "A recognised the White Rabbit: it was taking in a hurried nervous She was waking 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 Rabbt began. Alice gave a ittle 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 hastiy, 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

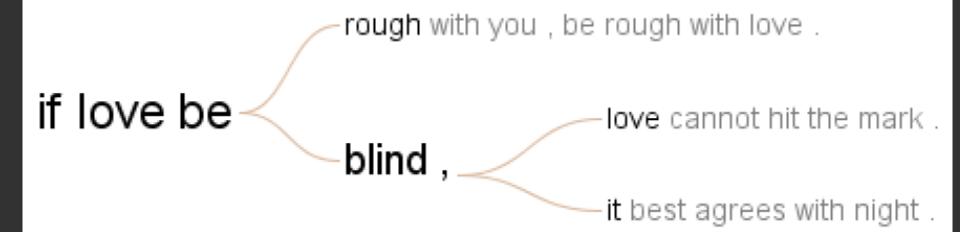
"Call the first witness," said the King; and the White Rabbit

6 X

Täitsbeped mad

pla

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.]

thygod

yourgod

 \nearrow thine heart , and with all thy soul , \checkmark

and with all thy might that thou mayest live .

thy heart , and with all thy soul , and with all thy $<\!\!<$

strength , an

mind

keep his charge , and his statutes , and his judgments , and his commandments , alway .
and to walk ever in his ways ; then shalt thou add three cities more for thee , beside these three : 19 : that thou mayest obey his voice , and that thou mayest cleave unto him : for he is thy life , and the to walk in his ways , and to keep his commandments and his statutes and his judgments , that thou mayest live and to serve him with all your heart and with all your soul , 11 : 14 that i will give you the rain of your land walk in all his ways , and to keep his commandments , and to cleave unto him , and to serve him with all your soul of the walk in all his ways , and to cleave unto him ; 11 : 23 then will the lord drive out all these nations from with all your heart and with all your soul .

all ye his saints : for the lord preserveth the faithful , and plentifully rewardeth the proud doer . hate evil : he preserveth the souls of his saints ; he delivereth them out of the hand of the wicked because he hath heard my voice and my supplications .

-name of the lord, to be his servants, every one that keepeth the sabbath from polluting it, and taketh hold of my covenant -good, and establish judgment in the gate: it may be that the lord god of hosts will be gracious unto the remnant of joseph -evil; who pluck off their skin from off them, and their flesh from off their bones; 3: 3 who also eat the -truth and peace.

with all

other ; or else he will hold to the one , and despise the other . ye cannot serve god and mammon .__

6 : 25 therefore i say unto y
 16 : 14 and the pharisees a

uppermost

lord

rooms at feasts, and the chief seats in the synagogues, 23: 7 and greetings in the markets, and to be called of seats in the synagogues, and greetings in the markets.

father

love the

hath bestowed upon us , that we should be called the sons of god : therefore the world knoweth us not , because it knew him

brotherhood .

world , the love of the father is not in him .

brethren .

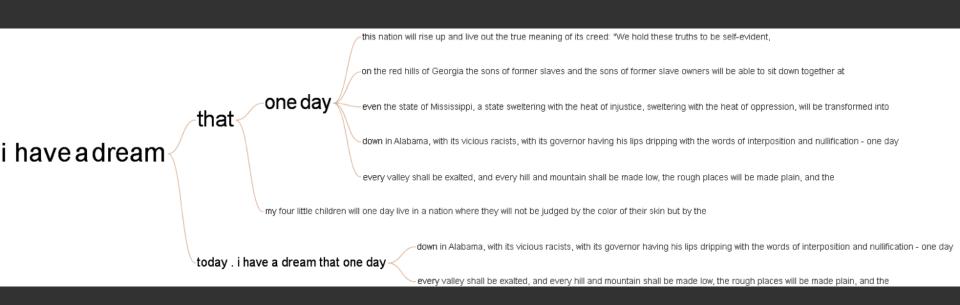
children of god , when we love god , and keep his commandments .

and as the father gave me commandment, even so i do

Filter Infrequent Runs

64,028 bing 64,028 king hito and a hits 111 earth------lordday 004 and landlord 10 bed . chidree insite -----ID8 1 the tana tan the hoame 1001 -land-12.00 sone . bund children of Israel neonle of son -APR/S term bott house--uf and and annel anne

Recurrent Themes in Speeches



search

 \mathbf{v}

many eyes

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

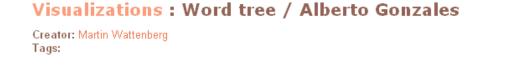
legal terms of use

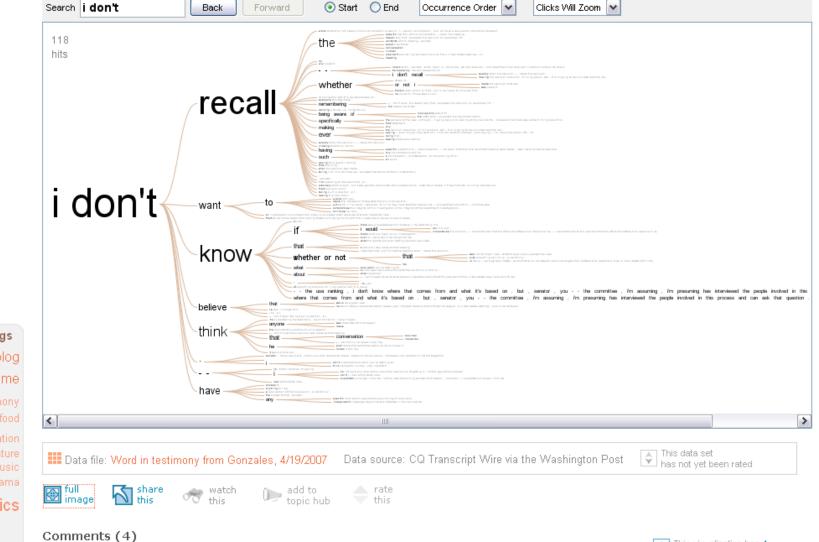
Popular Dataset Tags 2007 2008 bible blog books CENSUS crime education eharmony election energy food health inauguration internet ireland literature lyrics media music network obama

people politics population

president prices religion

currently showing





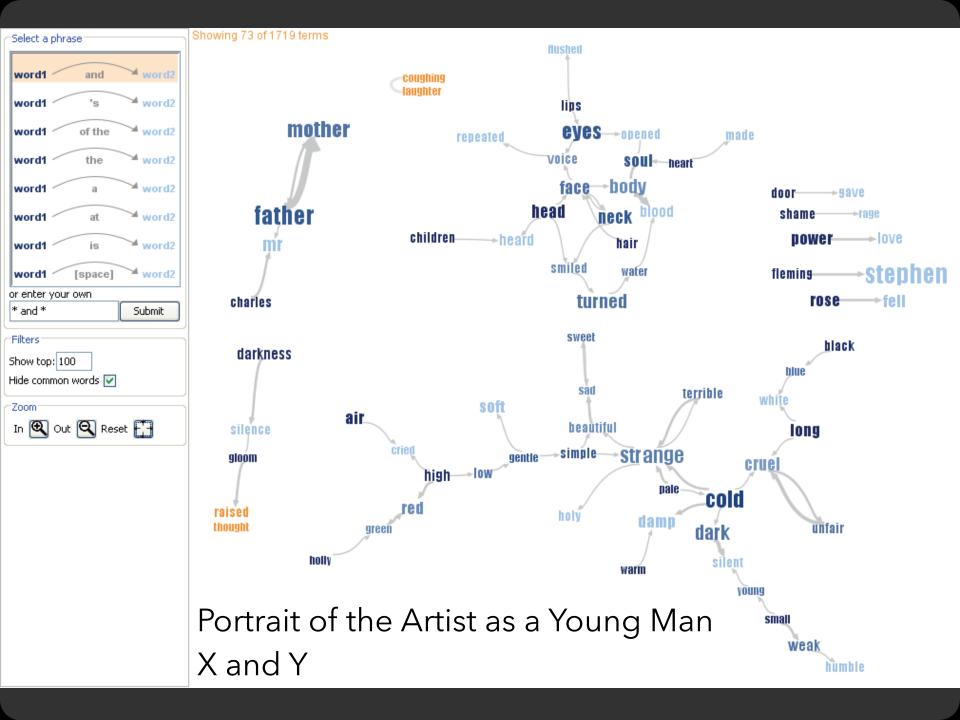
Glimpses of Structure...

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

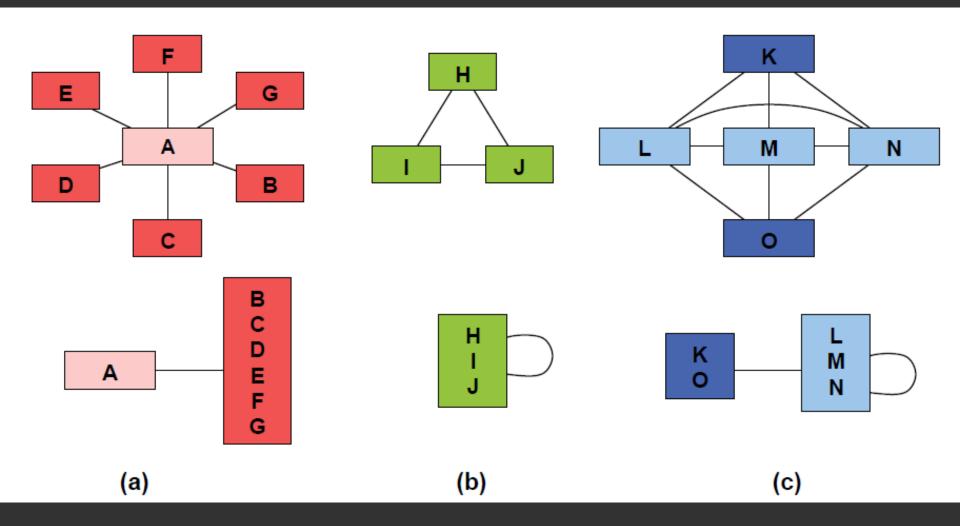
Lexical:<A> at Syntactic:<Noun> <Verb> <Object>

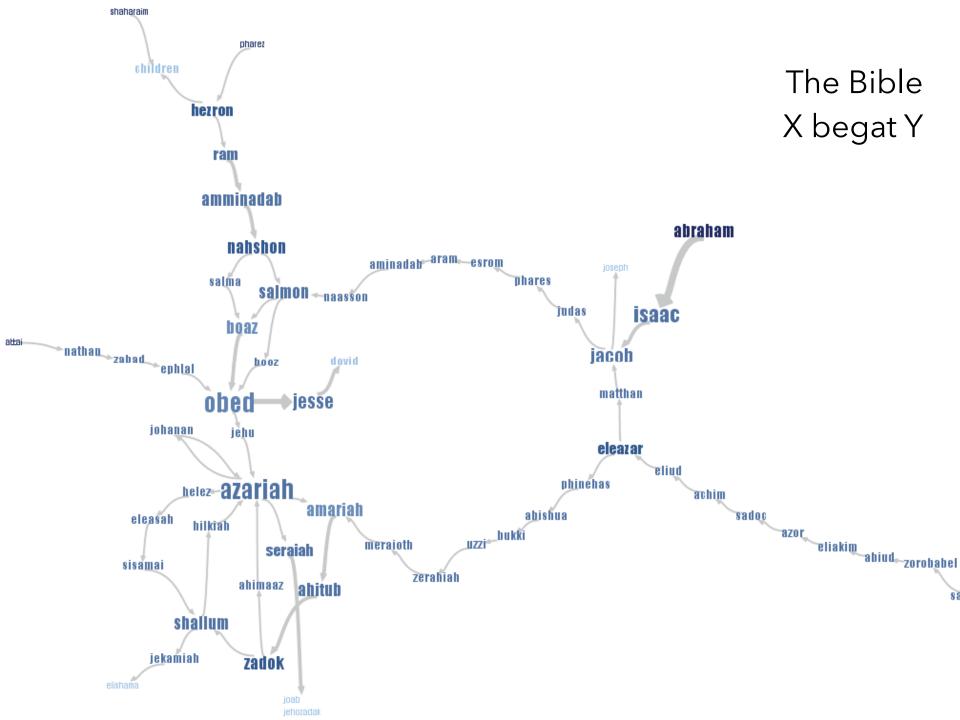
Phrase Nets [van Ham et al.]

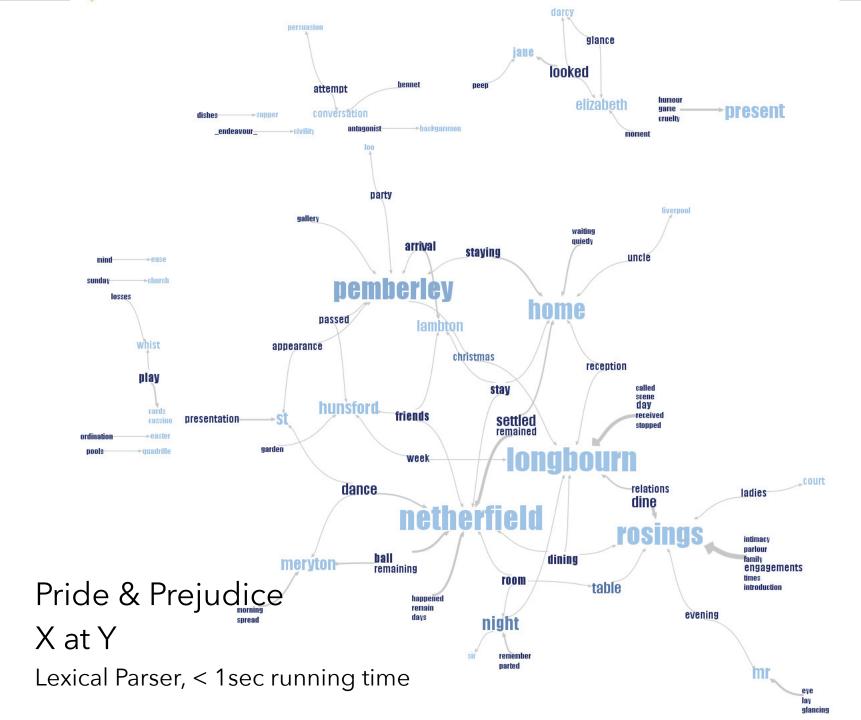
Look for specific **linking patterns** in the text: 'A and B', 'A at B', 'A of B', etc Could be output of regexp or parser. Visualize patterns in a node-link view Occurrences -> Node size Pattern position -> Edge direction

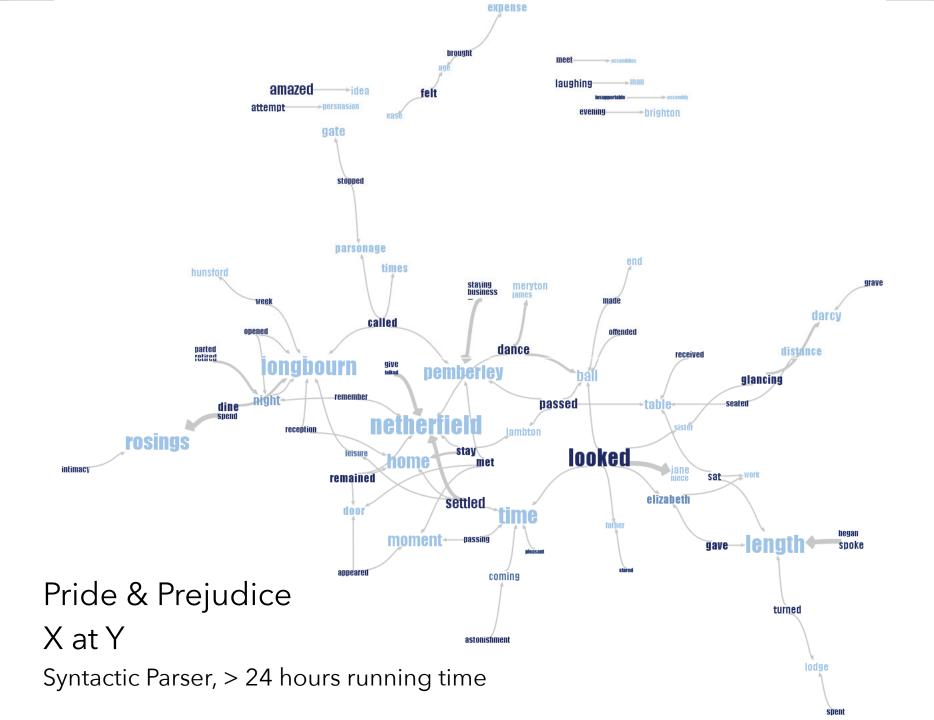


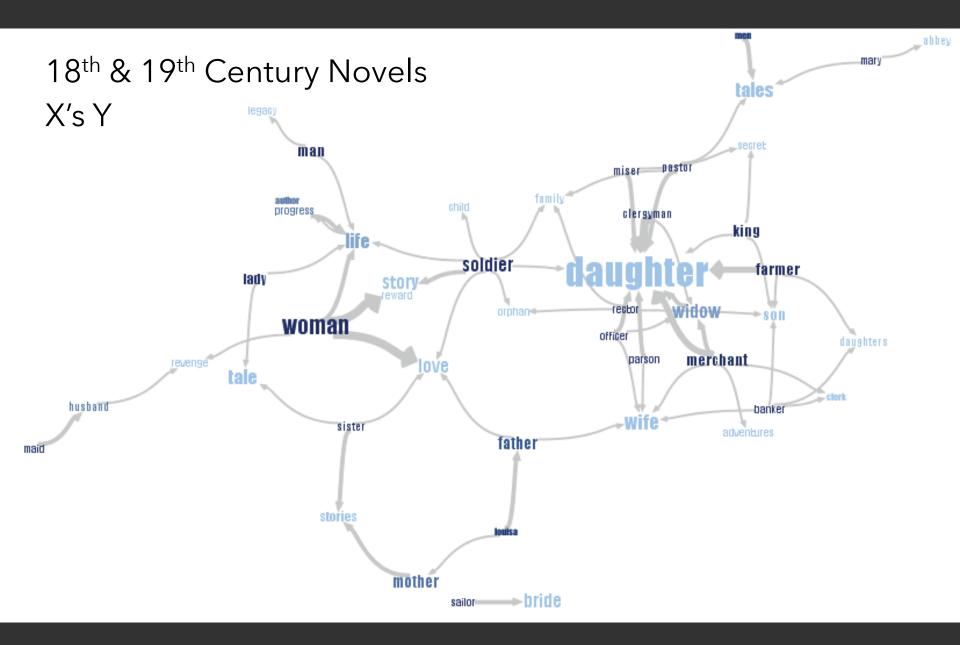
Node Grouping

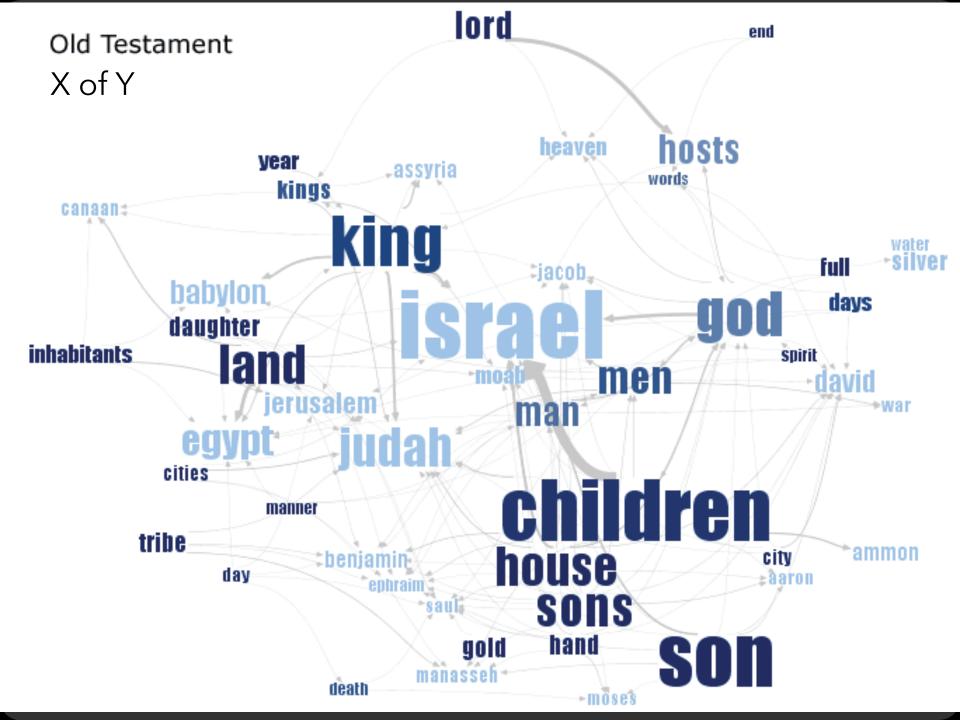












.john New Testament mother X of Y days city blood nazareth paul galilee esus epistle dàvid gospel sight sons grace truthword spirit SO full christ nower death⁴ faith day life glory love wisdom body house peace men light righteousness abraham judgment words man kingdom law manner wrath people hand heaven children **Fisrael** king things angels siñ moses judaea land

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

Evolving Documents

	Tool	ls T <u>a</u> ble <u>W</u> indow <u>H</u> elp Ado	be PDF Acrobat Comments	
	ABC	Spelling and Grammar F7	100% 👻 🙉 其 Keau 🔤 🛃 Normal 📼	
	í,	<u>R</u> esearch Alt+Click	pw + 😥 🎲 🖓 + 🏡 + 🛄 🕸 + 🎲 📄 📘	
		Language •		
		Word Count		
		Speech	· · · 1 · · · 1 · · · 2 · · · 1 · · · 3 · · · 1	
		Shared Wor <u>k</u> space		
	12	<u>Track Changes</u> Ctrl+Shift+E		
		Letters and Mailings	Track Changes Icons appear once you	
l		Customize	select "Track Changes" from the "Tools	
		Options	Menu"	

This is a test document to demonstrate the use of tracking changes. The characters in black font represent the original document while the characters in red font represent the changes which are being tracked



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) . . <u>12:01, 20 Aug 2003</u> . . <u>Dysprosia</u> (neaten to do, rearrange see also)
- (cur) (last) . . <u>11:59, 20 Aug 2003</u> . . <u>Patrick</u>
- (cur) (last) . . . <u>11:52, 20 Aug 2003</u> . . . <u>81.203.98.109</u>
- (cur) (last) . . M <u>18:36, 6 Aug 2003</u> . . <u>Manika</u> (corrected spelling)
- (cur) (last) . . <u>18:32, 6 Aug 2003</u> . . <u>Daniel Quinlan</u> (removing obscure heraldry information, belongs on [[heraldry]] if anywhere)
- (cur) (last) . . 15:21, 6 Aug 2003 . . Rmhermen
- (cur) (last) . . <u>15:08, 6 Aug 2003</u> . . <u>Cyp</u> (Chocolate often has odd shapes.)
- (cur) (last) . . <u>19:14, 3 Aug 2003</u> . . <u>Daniel C. Boyer</u> ("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/

😂 Java - WelcomePageDispatchAction. java - Eclipse SDK

File Edit Navigate Search Project Run Window Help

i 📬 ▾ 🔚 🗁 i 췒 ▾ 🏇 ▾ 🖸 ▾ 🏰 📽 🞯 ▾ i 🕭 🖋 ▾ i ½ ፦ 🖓 ▾ i ½ ፦ 🖓 ▾ ⇔ ⇔ ↔ ↔ 🖆 👔 🖓 ▾ ii ½ - 🖏					
🛱 Package Explorer 🛿 🍃 Hierarchy 🔲 🔄 🌄 🗖 🗖	Fragment Comparison 🛛 🚺 LoginDispatchAction.	CollectInformationDi			
🖃 🔂 Struts demo 🔼	Previous Version	Current Version			
- 🗁 src	// End of user imports	// End of user imports			
🖨 🥵 WEB-INF/src					
com.mia_software.booster.page_flow_example tom.mia_software.booster.page_flow_example	public class WelcomePageDispatchAction	public class WelcomePageDispatchAct			
🚊 🖶 com.mia_software.booster.page_flow_example	// associated forward definitions	// associated forward definitio			
🕀 🗓 CollectInformationDispatchAction.java	public final static String COLLECT	public final static String COLL			
🕀 🛄 CollectInformationForm.java	public final static String LOGOUTAN	public final static String LOGO			
😟 🖳 ResultDispatchAction.java	public final static String TEST1 To	public final static String TEST			
🕀 🗓 ResultForm.java					
🕀 🗓 Test 1 Dispatch Action . java	// inherited forward definitions	// inherited forward definition			
🕀 🛄 Test1Form.java					
😥 🛄 WelcomePageDispatchAction. java					
🗄 🛄 WelcomePageForm.java	// dispatch action methods declarat	// dispatch action methods decl			
🕀 🔠 🔠 com.mia_software.struts.generic.back	public ActionForward enter(ActionMa	public ActionForward enter(Acti			
🕀 🖶 com.mia_software.struts.generic.form 🛛 🗸	ActionForward actionForward = 1	ActionForward actionForward			
	if (form != null) {	if (form != null) {			
🞯 Generation Results 🛛 🔗 🗖 🗖	WelcomePageForm currentForm	WelcomePageForm current			
	// execute code on exit of	// execute code on exit			
📮 🝺 Results (30) 🔼	currentForm.onExit();	currentForm.onExit();			
🖻 📝 ResultDispatchAction. java	int i_ENTERXXX=0;	CollectInformationForm			
🖻 📝 Test1DispatchAction.java	CollectInformationForm col.	1			
😑 💀 WelcomePageDispatchAction.java		// Start of user code :			
Generated Code	// Start of user code : ret				
Manual Code					
Manu 🗸 Sort					
Gener 🗸 Group by Status					
Manu • File Name					
Genei Relative Path					
Manu Full Path					
CollectInf Compare with Previous Version					
CollectInf Show In					
! ⊡◆		i a 🔝 @ 😣 🚍 🛷			

FR Français (France) [💈 📮

^

Diff style: Side-by-side 🗸

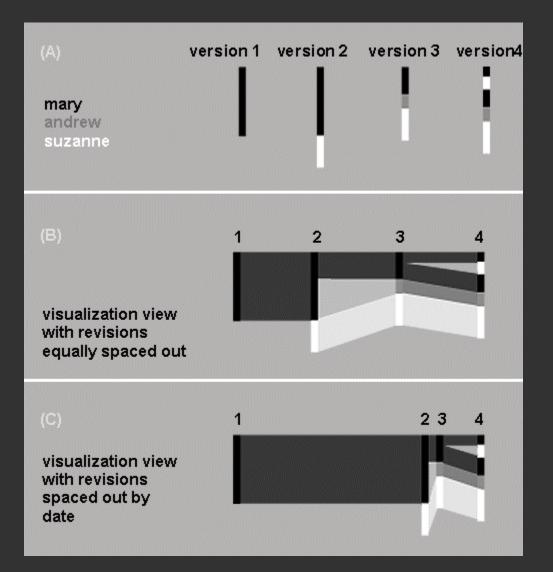
Enable syntax coloring

Files Changed:

1. sshconsole.js: 1 change [1]

/home/toddw/src/sshconsole-read-only/content/sshconsole.js								
	50 lines hidden [Expand]							
51	_term = new VT100(80, 24, "term");	51	_term = new VT100(80, 24, "term");					
52	//_term.debug_ = 1;	52	<pre>//_term.debug_ = 1;</pre>					
53	_term.curs_set(true, true, _term_box_element);	53	_term.curs_set(true, true, _term_box_element);					
54	_term.noecho();	54	_term.noecho();					
55		55						
56	<pre>// Replace the go_getch_ function with our own, this is called</pre>	56	<pre>// Replace the go_getch_ function with our own, this is called</pre>					
57	<pre>// for every keypress that is passed through the terminal to the</pre>	57	<pre>// for every keypress that is passed through the terminal to the</pre>					
58	<pre>// remote server. The character is already converted into the</pre>	58	<pre>// remote server. The character is already converted into the</pre>					
59	<pre>// required VT100 character sequence(s).</pre>	59	<pre>// required VT100 character sequence(s).</pre>					
60	<pre>VT100.go_getch_ = function() {</pre>	60	<pre>VT100.go_getch_ = function() {</pre>					
61	<pre>var vt = VT100.the_vt_;</pre>	61	<pre>var vt = VT100.the_vt_;</pre>					
62	if (vt === <mark>undefined</mark>) {	62	if (vt === <mark>somevalue</mark>) {					
63	return;	63	return;					
64	}	64	}					
65	<pre>var ch = vt.key_bufshift();</pre>	65	<pre>var ch = vt.key_bufshift();</pre>					
66	//dump("go_getch_:: ch: '" + ch + "'\n");							
67	if (ch === undefined) {	66	if (ch === undefined) {					
68	return;	67	return;					
69	}	68	}					
70	if (vt.echo_ && ch.length == 1) {	69	if (vt.echo_ & ch.length == 1) {					
71	vt.addch(ch);	70	vt.addch(ch);					
		71	vt.refres();					
72	}	72	}					
73	if (_ssh_channel) {	73	if (_ssh_channel) {					
74	_ssh_channel.sendStdin(ch);	74	_ssh_channel.sendStdin(ch);					
75	}	75	}					
76	}	76	}					
77		77						
78	<pre>var serverTextbox = document.getElementById("sshconsole_server_textbox");</pre>	78	<pre>var serverTextbox = document.getElementById("sshconsole_server_textbox");</pre>					
79	var connectionText;	79	var connectionText;					
80	if ('connectionText' in window.arguments[0]) {	80	if ('connectionText' in window.arguments[0]) {					
81	<pre>connectionText = window.arguments[0].connectionText;</pre>	81	<pre>connectionText = window.arguments[0].connectionText;</pre>					
82	} else {	82	} else {					
	174 lines	hidden []	Expand					

History Flow [Viegas et al.]



"Abortion"

posts

from Wikipedia

Conversion script 1

authors

Zundark

The Epopt

Shubenstein

Maverie149

Theanthrope

Drehmword

Comembert

Hephoestos

MyRedDice

Kingturtie

COLOR Ne group

December

2001

🔚 individual 🛛 🔛 text changes 💥 text age

SPACING O date O versions

198.37.26.168 Wikipedia History Flow (IBM)

Abortion

(Revision as of 22:56 4 Jun 2003)

"Abortion," in its most commonly used se refers to the deliberate early termination pregnancy, resulting in the death of the gr fetus, [1] Medically, the term also refers t early termination of a pregnancy by natur ("spontaneous abortion" or miscarriage, w 1 in 5 of all pregnancies, usually within the weeks) or to the cessation of normal grow body part or organ. What follows is a disci the issues related to deliberate or "induceabortion.

Methods

Depending on the stage of pregnancy an a performed by a number of different metho a chemical abortion is the usual method, t mifepristone is usually the only legal meth although research has uncovered similar e from methotrexate and misoprostol. Conc with chemical abortion and extending up u around the fifteenth week suction-aspiration vacuum abortion is the most common app replacing the more risky dilation and curet C). From the fifteenth week up until aroun eighteenth week a surgical dilation and ex (D & E) is used.

As the fetus size increases other technique be used to secure abortion in the third trip premature expulsion of the fetus can be in with prostaglandin, this can be coupled with injecting the amniotic fluid with saline or u solution. Very late abortions can be broug by the controversal intact dilation and extr & X) or a hysterotomy abortion, similar to caesarian section.

The controversy

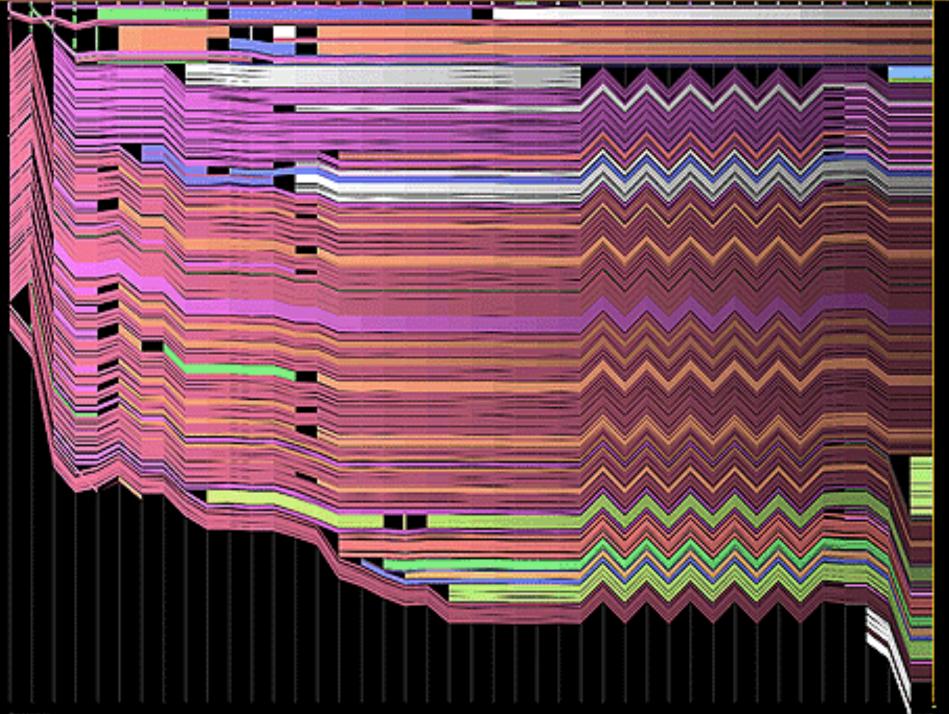
The morality and legality of abortion is a l important topic in applied ethics and is als discussed by legal scholars and religious p Important facts about abortion are also re by sociologists and historians.

Abortion has been common in most societ although it has often been opposed by sor institutionalized religions and governments century politics in the <u>United States</u> and <u>En</u> abortion became commonly accepted by the 20th century. Additionally, abortion is accepted in China. India and other populo countries. The Catholic Church remains o the procedure, however, and in other cour notably the United States and the (predom Catholic) Republic of Ireland, the controve extremely active, to the extent that even of the respective positions are subject to I debate. While those on both sides of the are generally peaceful, if heated, in their i of their positions, the debate is sometimes characterized by violence. Though true of sides, this is more marked on the side of t opposed to abortion, because of what they the gravity and urgency of their views.

The central question

2003

The central question in the abortion debat clash of presumed or perceived rights. On hand, is a fetus (sometimes called the "un pro-life/anti-abortion advocates) a human with a right to life, and if so, at what point pregnancy does the fetus become human? other hand, is a fetus part of a woman's b



Conversations

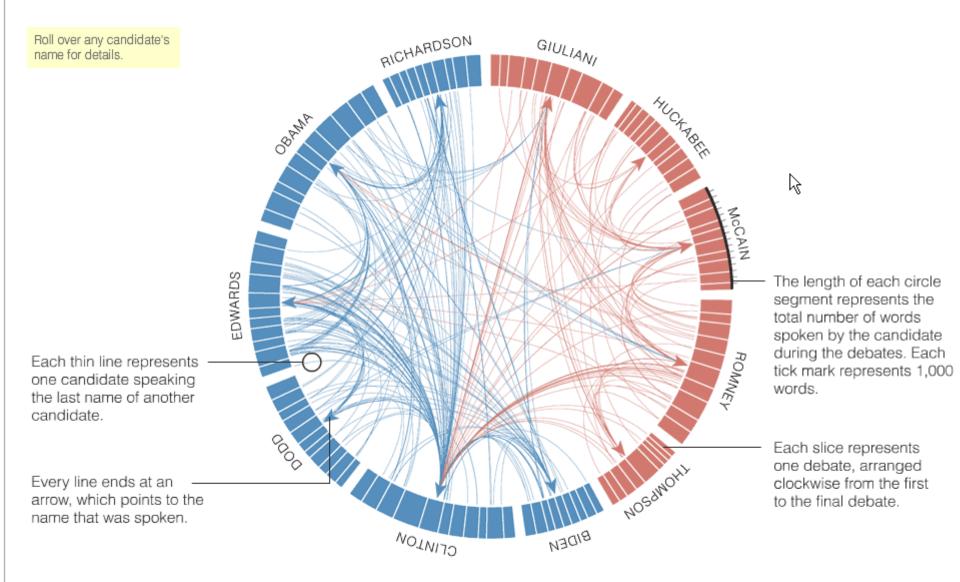
Visualizing Conversation

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.



Usenet Visualization [Viegas & Smith]

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



threads not initiated by author	threads initiated by author
	ەل •
	n
	Fel
	•••••••••••
	Ma
0	
	000000000000000000000000000000000000000
	000000000000000000000000000000000000000
0	
	Ap
	00000000000000
	vlay
	Ju
	n
	Au
·····	
	Se
	P
	00000000
	000000000000000000000000000000000000000

back to newsgroups

Week of Oct 21, 2001

......

0.....

8

subject / /ednesday Spooker ASF /ET #3 Anyone for breakfast	# of posts
/ednesday Spooker ASF	21
(ET #3 Anyone for breakfast	
unny Side Up ASF)	
unny Side Up ASF) sturday Ensemble and WET h nol Watch out! ASF hursday Combo-Post WET #	
h nol Watch outle ASF	
hursday Combo-Post V/ET #	
he Yellow Rose Inn A gitt to	
ET#1 JBP The First Time	
he Yellow Rose Inn A gift to IET #1 JBP The First Time /e Love the Earth ASF Ionday Spooker "The Sight"	16
onday Spooker "The Sight"	16
heberge "Le Vent Se Leve"	14
oliday Ton #1 1	12
pooker du Jour) eginning ASF Short and econd Try A Kalie for Suzy	13
eginning ASE Short and	
econd Try A Kalie for Suzy ome On a Safari With Me uesday Spooker ASF urses, Folled Again ASF	12
ome On a Safari With Me	11
uesday Spooker ASF	11
urses, Foiled Again ASF	10
lalloween Togs Take Two) eauty of the Fury Jim Warren hought I saw ? ASF rednesday Evening at the Con econd Try A Kalle for Suzy	
eauty of the Fury Jim Warren	
hought I saw ? ASF	
rednesday Evening at the Con	
econd Try A Kalie for Suzy	
rank Was A Monster ASF	

-

subject	# of posts
Sunday Twofer ASF ()	9
Chopsticks/A Jilly fake	
Oh not Trouble in Discworld!	7
WET your thirst I ASP	.6
A pretty for you. Reposted fro	5
Saturday Spooker ASF	5
Sample Previous install Upgr	4 - 1
Tennessee weather tonite	4
WET - Well I am not smiling!	4
Somethin' mushy <ast></ast>	3 (
Getting seasonal with workin	3
A Haunted House)	3
do you wonder what debits be	3
Question: Ethics of posters in	3
For Jerry	3
Olu's Tribe slightly rated	3
WET - Glass Bottles	3
Peace Train <asf></asf>	2 1
Arrival at Stewart Island II	2
WET 195 Wrap-up	2
Cat O'Lantern	2
Put a Spell on You (Happy H.,	2
Goodbye to Summer - A Timel.	2
Two Pumpkins In A Strange B	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
Still Heading South II	2
WET- Frank Sinatra - The Man	2
WETAutumn	2
Purple Martin ASF	2 2 2 2 2 2
Opposites Attract	2
Time	2
	× ×

1

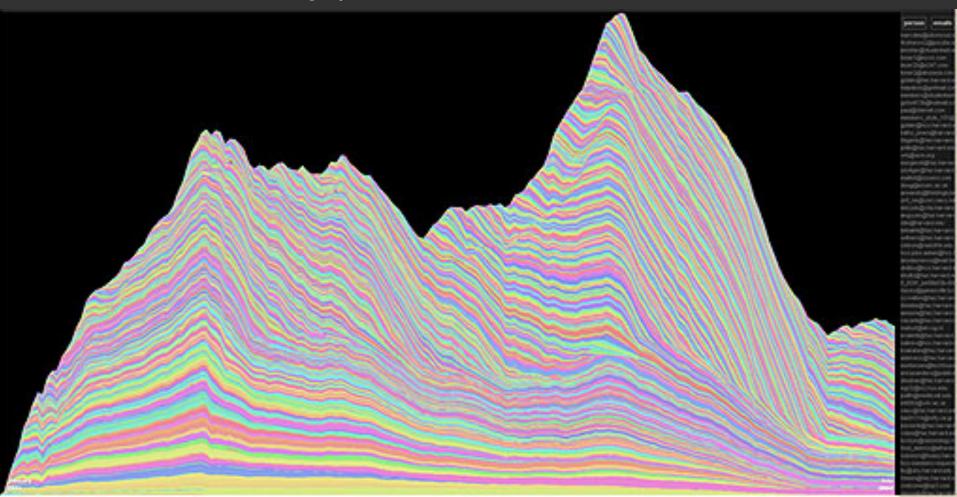
author:

jillyb@mail.com

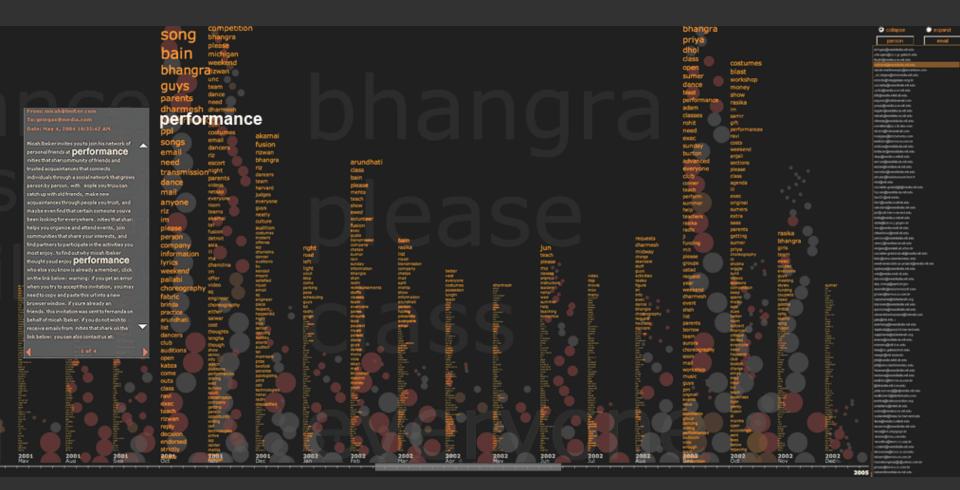


Email Mountain [Viegas]

Conversation by person over time (who x when).



Themail [Viegas]

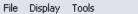


One person over time, TF.IDF weighted terms

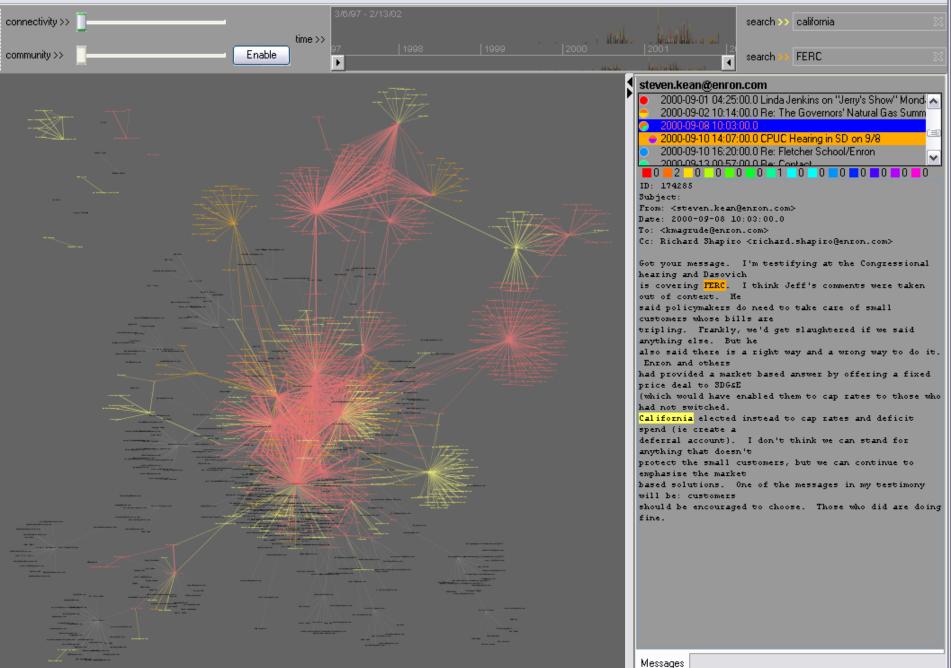
Enron E-Mail Corpus

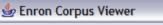
[Heer]

👙 Enron Corpus Viewer

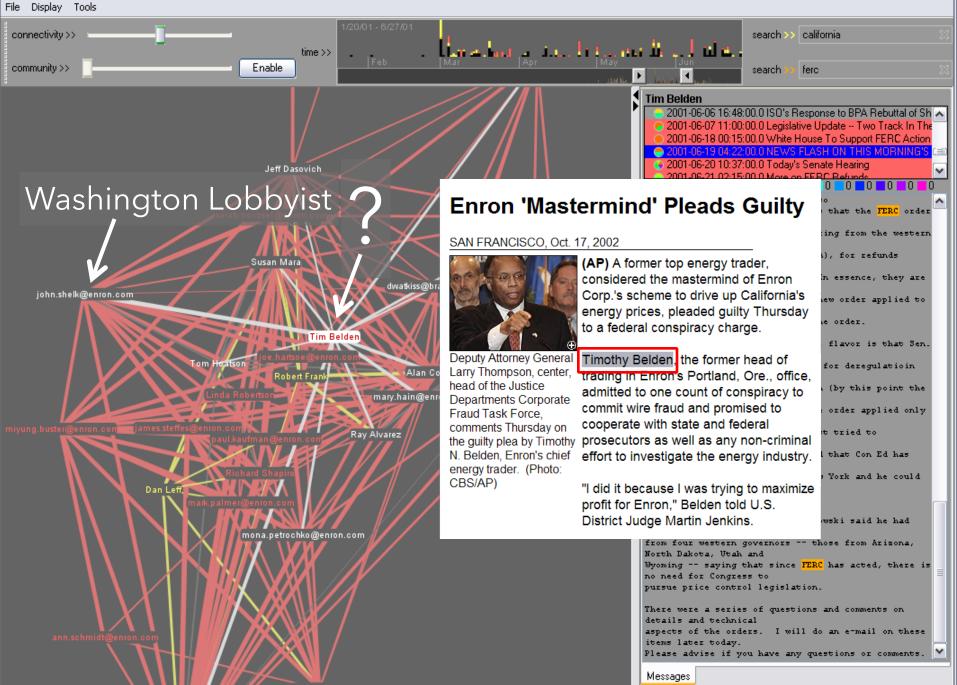












Document Collections

+ about + permalink	+ SELECT ALL COUN	TRIES AUSTRALIA	AUSTRIA	CANADA	FRANCE	DEUTSCHLAND	INDIA	ITALIA	NEW ZEALAND	ESPANA	U.K .	U.S.
Pakist hopes			Clinton criticise Israeli	S	Cricket Pakista			Was Fallon funny?	CONCERT REVIEW: Britney Yet to Hit Stride	Spotlight still follows the Jonas Brothers	The Rush To Bash Rush	The Strengths and Weaknesses of the Chandra Levy Case
intern			demolit ns in	io	fled du	ring		How Could Rihanna Take Back Chris Brown?	Jason & Molly: 'Fighting Like Hell to Make It Work'	tickets are hot: Have you bought yours yet?	Marines: Celebratin victory, Multiple villaraigos errors caused San Diego	Stiles a found ahead guilty on all counts
arena			E.Jerus em	al	ambus	Π		Economy to Dominate Annual Chinese Gathering	Octo gets -Home 8 cast Times: Report Mile Mov	ical' stresses a new of reading to Boone students Party Wips to Warter, Barainey Vie Barainey Marty digetors	Crash Spending bil Obama: billons in Contractin earmarks g overhaul	Supreme Coart Divided Una Case Tapped to Head FCC
British	Command er says	World court issues arrest	Afghan Bor Election Thr		Coast Guard	Man ny, LA	Girardi says hip bothered	Michael Jackson seeks comeback in London Oil Gains a		i Moli ² i partine inter i statistica i care i care i statistica i care i c	to save Emand's likely replacement \$40 billion no one cares a year	Galance is for the second seco
Prime Minister to Address US	Iran missiles can reach Israel	warrant for Sudan's Bashir	Says Early Ser Vote Not in S Possible Afg	nadian vicemen Southern hanistan _{Gandhi kin}	Calls Off Search for	CIOSE Suns- Be	A-Rod last year ackham's Wednesd	Second Day of Speculation China Will Bo Stimulus	on spends \$ per car o	n US Mac de s, compu ⁻)	ces new TSMC: What asktop are they ters thinking?	Push to Delivery of boost FDA Healthcare food safety Reform Bil system AUDs affecting over 50s: WHO disgnosed with cancer
Congress	atom sites China, US	Has Pakistan become the central	Africa: troops enter United States May Boycott UN Racism	appeals for return of memorabil ia	Missing Players,	Preview ex tru	afty new an deal poses the ith	AIG Blast May	ollywood mulis Costco fe without increas lockbuster discou	profit falls urt by ed	CEBIT-IT Phew! industry will Asteroid's be back, passing Schwarzen was a egger says cosmic near-miss	NC doctor Concerted helps Effort Needed colleagues to Fight Drug- track online Resistant Flu reviews Strain
Interim Leader Takes Over in Guinea-Bissau	opposed to NKorean missile launch:	front? Zimbabwe' s Tsvangirai calls for end to sanctions	Germany: 2 still missing after building collapse	lan hoping pr 'world's est job' menoan: Gov't, urch to Stare Cost Pope's Visit nnedy to receive norary knighthood m Britain	Fellow Boater	Trail n Blazers Sta Preview Red Wings- LA	Navens Drew is given back to the second to t	Concentration	BS Gets Constant of the second	Square Root Day revelers t are start for a start stat sta	Conserveition de la conserveitinte de la conserveition de la conserveition de la	Market bergenergiesen and Statigenergiesen bergesengenergiesen werdenergiesengenergiesen werdenergiesengenergiesen werdenergiesengenergiesen werdenergiesengenergiesen werdenergiesengenergiesengenergiesen werdenergieseng
Wednesday March 4, 200	envoy 19 8:09		plane crash					Payment	1		Again LAYOUT: •	squarified standard
ARCHIVED THU FRI SAT SUN MON	VEST. TODAY HOW					LESS TH	HAN 10 MINUTES AGO 🔽 🗰	ORLD MATION	V BUSINESS	TECHNOLOGY	SPORTS ENTERTI	NIMMENT 🗸 HEALTH
00:00 06:00 12:00 18:00						MORE TH MOR	Han 10 minutes ago Re than 1 hour ago					

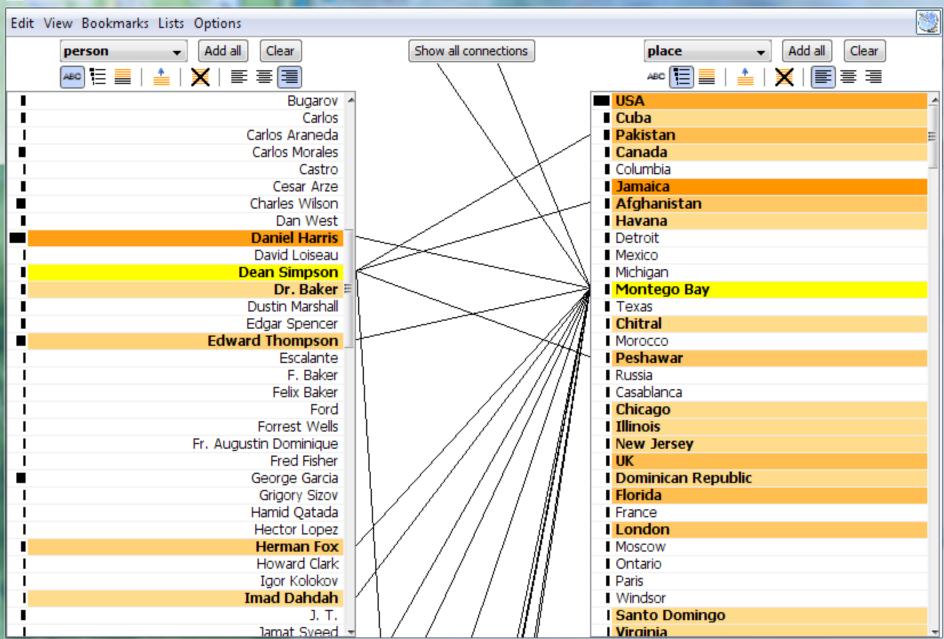
NewsMap: Google News Treemap [Weskamp]

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



х

Similarity & Clustering

Compute vector distance among docs For TF.IDF, typically cosine distance Similarity measure can be used to cluster

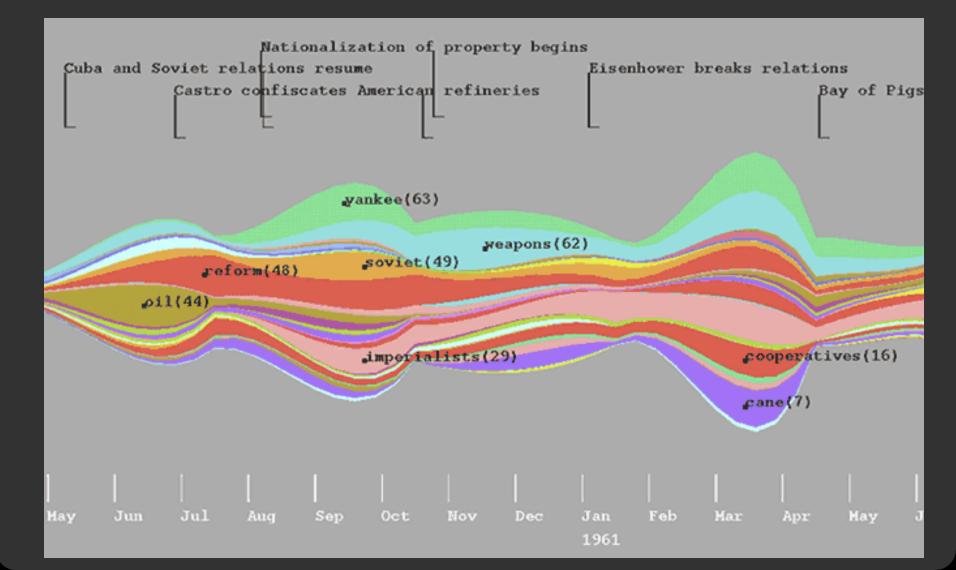
Topic modeling

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

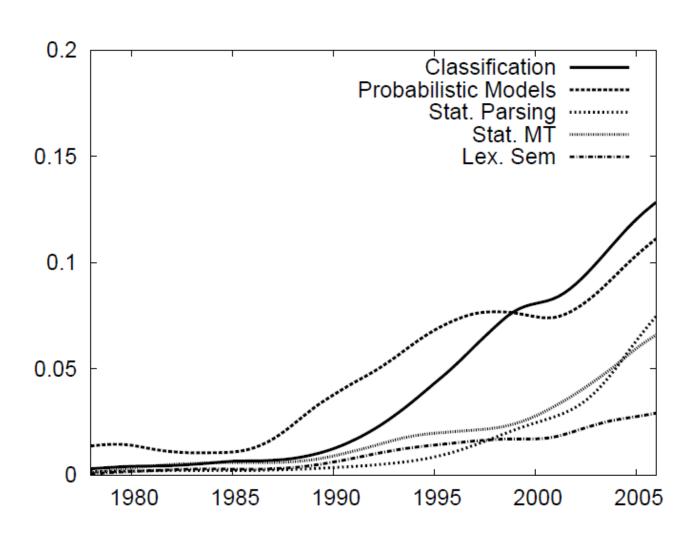
Parallel Tag Clouds [Collins et al.]

afficiavit adjourned abuse appeal ballo/ accused about abuse adverted allocatur adequate agency bankruptcy bargaining argument banc agency alia alia argued analysis aliens affirmed benefit affirmed assistant brief asked agency's anent black annuity barge allocution antitrust appropriate attached aid capital coal appellee boat antidumping appellant app called arbitration cargo binding authority asylum candidate arbitration brief application charter asbestos argued appeal appellant's bargaining broker asbestos court case art cited circuit closure brief assets coverag appellee argument conspiracy board commenced bankruptcy cited collateral broadcast defendant 100.00 class damages because complaint crack cable asseveration claim believe contended before defendant copy commerce capricious defendant's denied compositio below court copyright benefit debtor fisciplinar carrier compounds boat coal disability drilling enough crime county comprising denied distribution dba competition bottlers construction determine brief disability declared court conspiracy fire district costs denial defendant disfavor contract death district contention gang class deportation data disenfranchised aas drug commonwealth disposition get doc discrimination habeas gun discretion defendant court's employees decision evidence emissions doctrine context foreign disenfranchised del had creditors homestead description farm disposition employees filed firearm estoppel fraud exemption debtor dozer device ensued harassing indemnity district decisional firearm grams examination explanatory error disclosed election have iniury exercise follows errs forthwith embodied facilities had event denied electors instant grievance help factfinding injunction except gas fiduciary equivalent furnished disclosed her ferritin insurance immunities hazard her inter hereby further fear have dispensed guidelines gas interpretation his inequitable his him fish jack. insurance interna gnicing here intervenor here distribution jurists impair ivory infringement judgment habitat incarcerative law his keeplock district labor inasmuch job inmates jail hardship inference liability license insurance judgment iudicata invalid drug / job marks jury his inter error law lien interest material jury fact magistrate judge immigration invention memoranda loan migrant medical millions jurisdiction magistrate's limned urisdiction marihuana nevertheless from inventor mitigation petitioner marijuana kilogram methamphetamine his land lst legislation maritime nonstatutory layer interlocutory pipelines opinion lawyer months may might ioined mitigation motion liability ordinance oral preceding more might office plaintiff legal nativo means negligence office mortgage payday promulgated opinion majority more novo order lung plaintiff's merchandise nre panel proposed market phase magistrate plausible one pain method principal paupers persuasive ostrich panel eceding oil postconviction point notes out material noninfringement rate qualified plaintiff quotation parish persecution para quantity plaintiff's rescript merits our said regulations racketeering miner's plaintiff's precedential police racial patent petition parent platform reversed say reinsurance pension prisoner mining rehearing record search patentee policy pneumoconiosis section plaintiff respect prisoner sav see reprinted opinion police sentence provided res sentence recovery security she product some plan pulmonary oral ref'd sexual public submittee rulemaking suggested protein sheriff see pursuant plenary suit suspended order pursuant supra removed she section reissue shareholders retard newiew students supra recommendation policy retirement see shares pneumoconiosis think rigging subd trial specie tit tentative service sterile renduct search said seaman turtle tit present provision suitable stock than unanimous servitude tusks shipper testified sentence simal recognized subway vesse thought town process stated skill sitting testimony unfavorable tribal summation specific trialworthy section published told told vote tariff suit summary tribe trial pulmonary unpublished structure trade unanimous vessel settlement usury voters transmission tribal want unanimous surface vacated union syrup vessel value viz sertence union what white under verdict water vehicle upon vaccination waybill whom which writ wrote vol warrant work zone veterans waste where would without would First Seventh DC Second Third Fourth Fifth Sixth Eighth Ninth Tenth Eleventh Federal

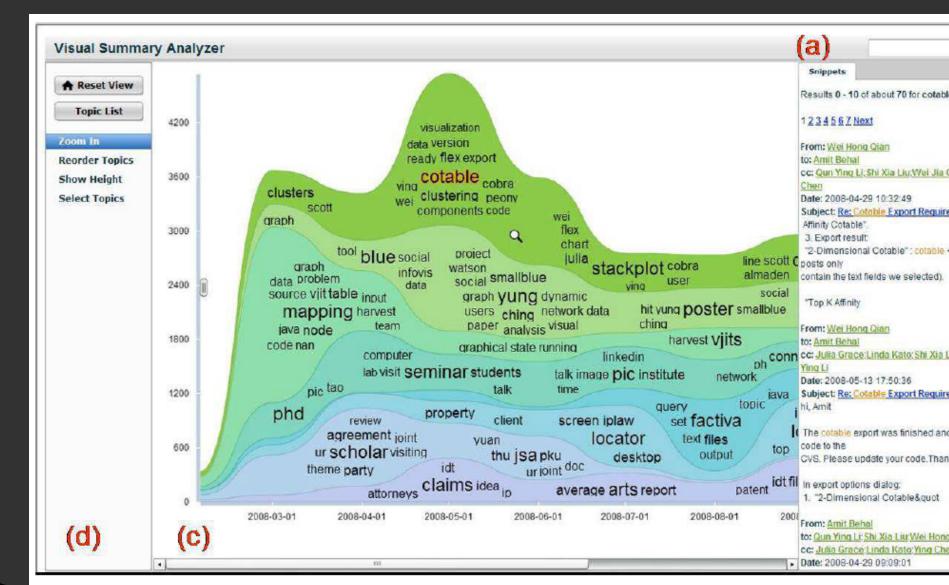
Theme River [Havre et al.]

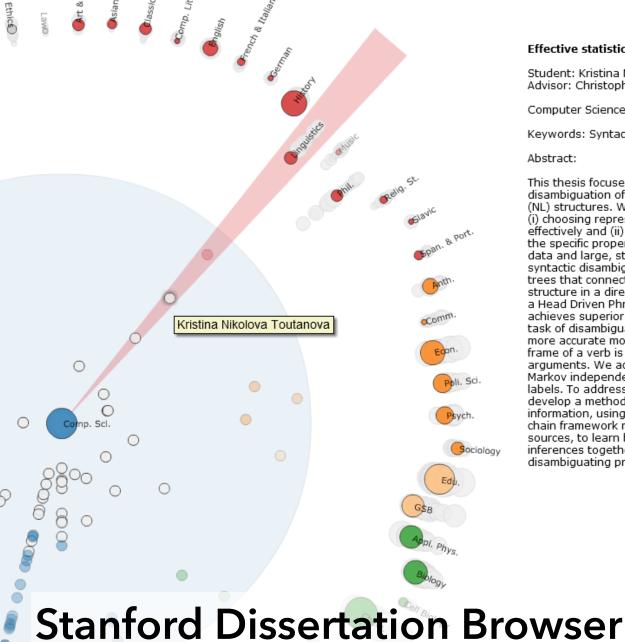


History of Comp. Ling. [Hall et al.]



Tiara [Wei et al.]





Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova Advisor: Christopher D. Manning

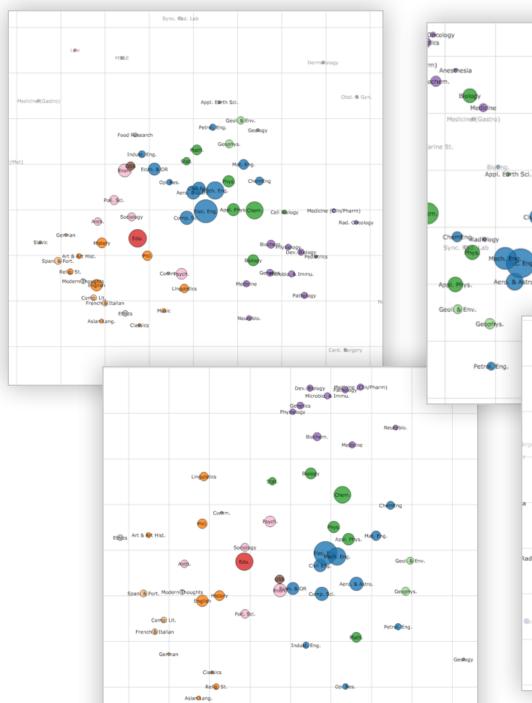
Computer Science (2005)

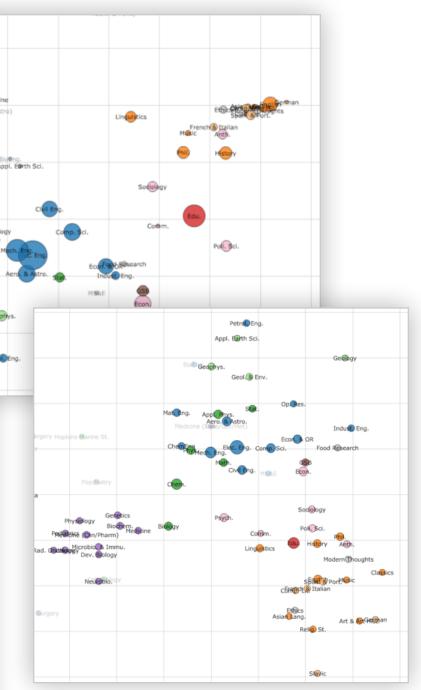
Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

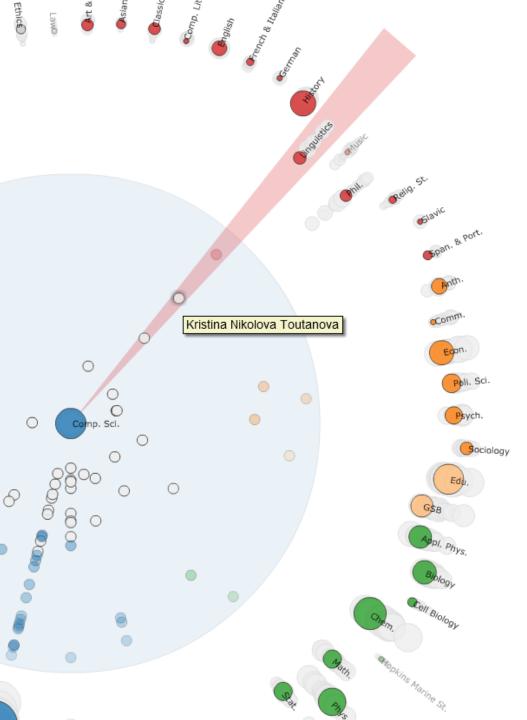
This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing 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.

with Jason Chuang, Dan Ramage & Christopher Manning









Effective statistical models for syntactic and semantic disambiguation

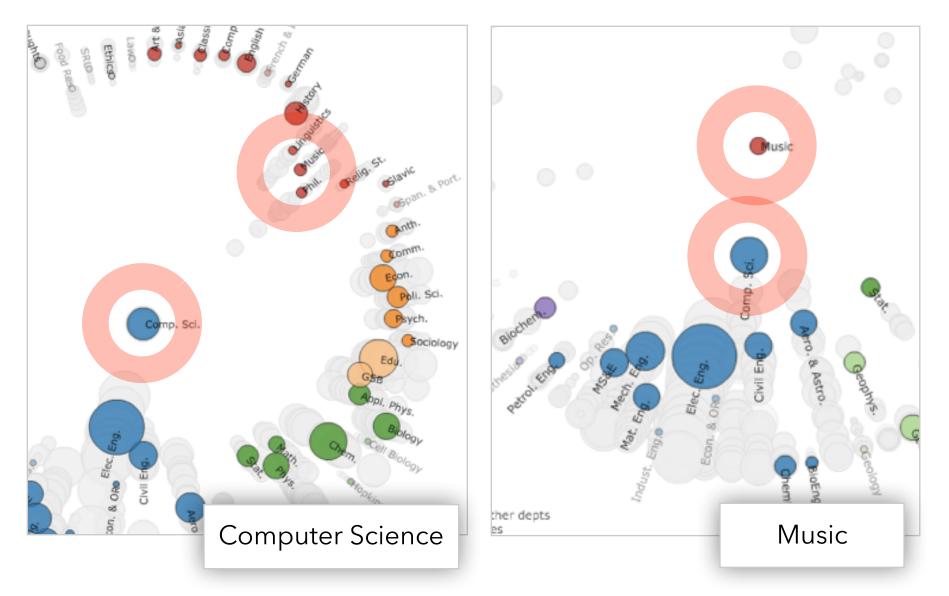
Student: Kristina Nikolova Toutanova Advisor: Christopher D. Manning

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.



"Word Borrowing" via Labeled LDA

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

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