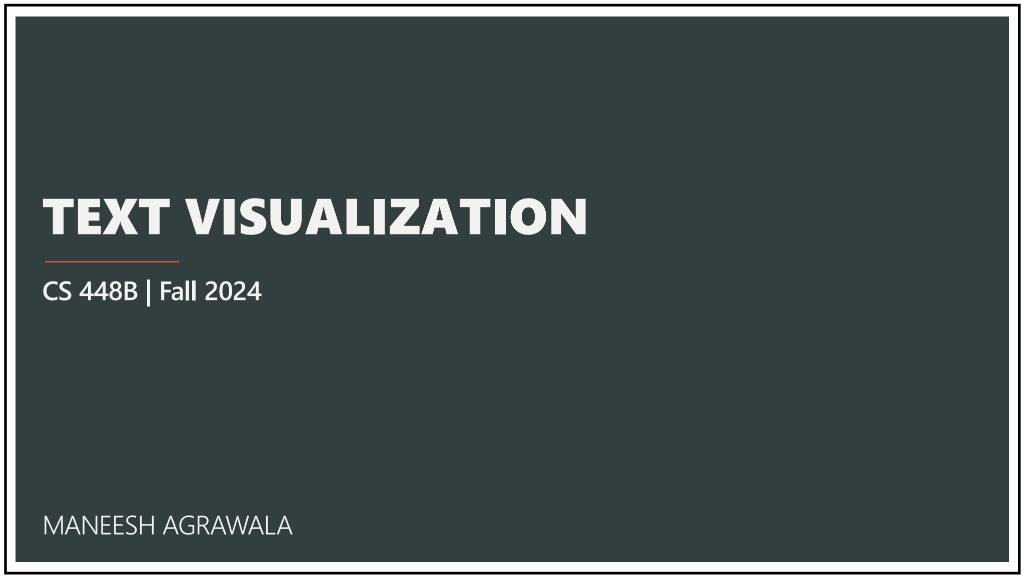
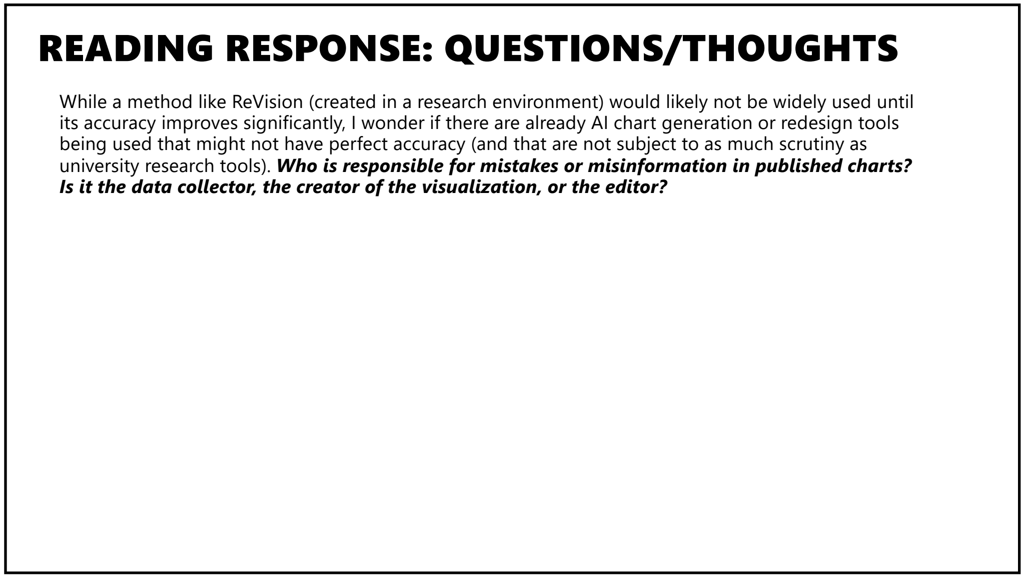


TEXT VISUALIZATION

CS 448B | Fall 2024

MANEESH AGRAWALA

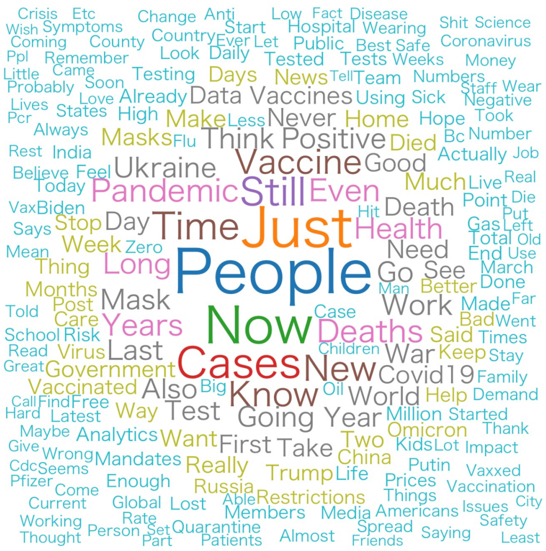
1



READING RESPONSE: QUESTIONS/THOUGHTS

While a method like ReVision (created in a research environment) would likely not be widely used until its accuracy improves significantly, I wonder if there are already AI chart generation or redesign tools being used that might not have perfect accuracy (and that are not subject to as much scrutiny as university research tools). ***Who is responsible for mistakes or misinformation in published charts? Is it the data collector, the creator of the visualization, or the editor?***

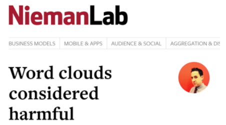
2



the daily report designing with web standards my gla

19 April 2005 12 noon edt

Tag clouds are the new mullets


NiemanLab

BUSINESS MODELS MOBILE & APPS AUDIENCE & SOCIAL AGGREGATION & DI

Word clouds considered harmful

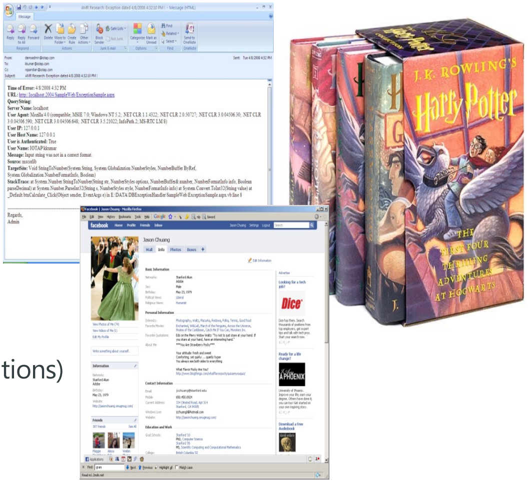
SINGLE DOCUMENTS AND COLLECTIONS

Documents

- Articles, books, novels
- Computer programs
- Email, Web pages, blogs
- Tags, comments

Collections of documents

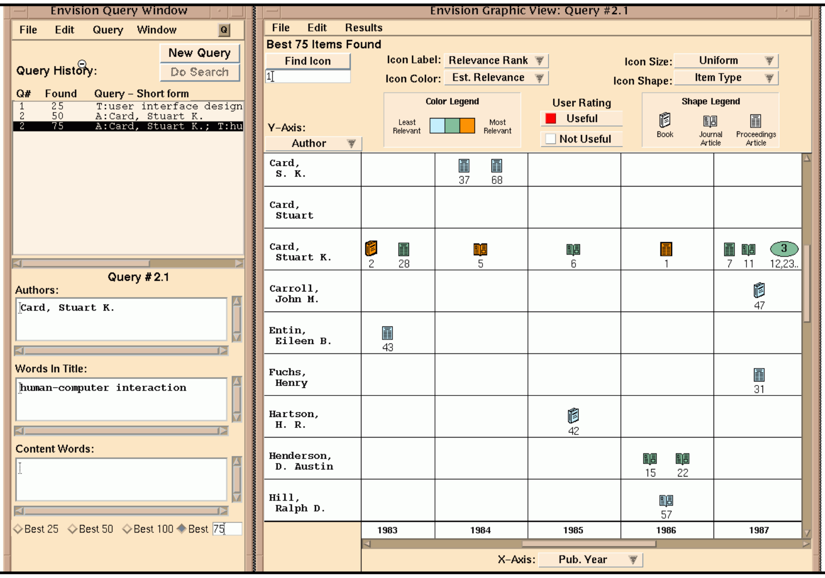
- Messages (e-mail, etc.)
- Social nets (profiles)
- Academic collaborations (publications)



TASKS

- What documents contain text about X?
- Which documents are of interest to me?
- Are there documents similar to this one?
- How are different words used in the collection?
- What are the main themes?
- How are themes distributed?
- What unifies/connects documents in a collection?

5



Envision Query Window

Query History:

Q#	Found	Query – Short form
1	25	T:user interface design
2	50	A:Card, Stuart K.
2	75	A:Card, Stuart K.; T:hu

Query # 2.1

Authors: Card, Stuart K.

Words In Title: human-computer interaction

Content Words:

◇ Best 25 ◇ Best 50 ◇ Best 100 ◆ Best 75

Envision Graphic View: Query #2.1

Best 75 Items Found

Icon Label: Relevance Rank

Icon Color: Est. Relevance

Icon Size: Uniform

Icon Shape: Item Type

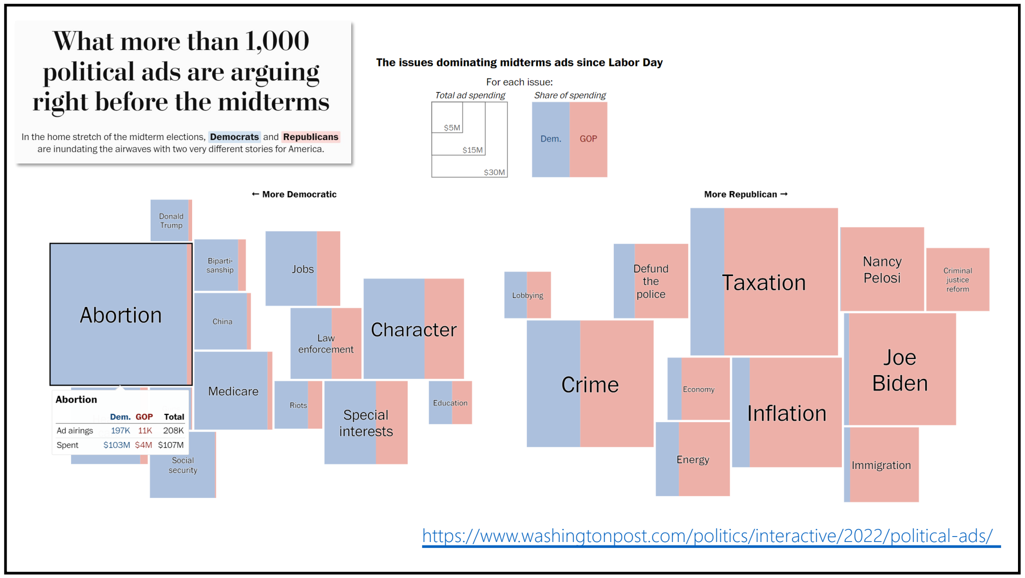
Y-Axis: Author

X-Axis: Pub. Year

	1983	1984	1985	1986	1987
Card, S. K.		37 68			
Card, Stuart					
Card, Stuart K.	2 28	5	6	1	7 11 3 12,23..
Carroll, John M.					47
Entin, Eileen B.	43				
Fuchs, Henry					31
Hartson, H. R.			42		
Henderson, D. Austin				15 22	
Hill, Ralph D.				57	

Fox et. al., Users, user interfaces, and objects: Envision, a digital library. *JASIS*, 44(8):480–491, 1993.

6



What more than 1,000 political ads are arguing right before the midterms

In the home stretch of the midterm elections, Democrats and Republicans are inundating the airwaves with two very different stories for America.

The issues dominating midterms ads since Labor Day

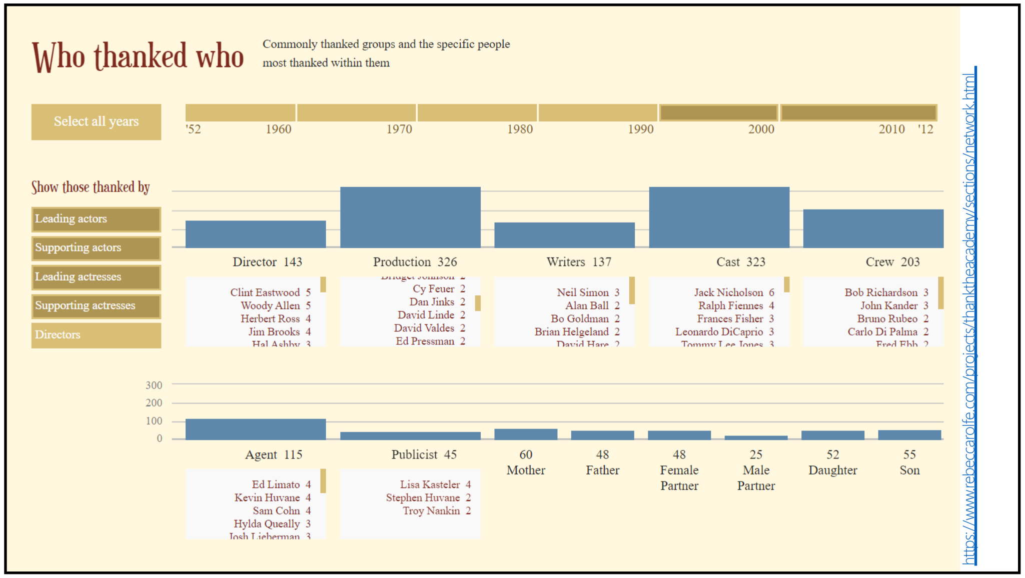
For each issue:

Abortion

	Dem.	GOP	Total
Ad airings	197K	11K	208K
Spent	$103M	$4M	$107M

https://www.washingtonpost.com/politics/interactive/2022/political-ads/

7



Who thanked who

Commonly thanked groups and the specific people most thanked within them

Select all years

Show those thanked by

- Leading actors
- Supporting actors
- Leading actresses
- Supporting actresses
- Directors

Director 143: Clint Eastwood 5, Woody Allen 5, Herbert Ross 4, Jim Brooks 4, Hal Ashby 3

Production 326: Bridget Johnson 2, Cy Feuer 2, Dan Jinks 2, David Linde 2, David Valdes 2, Ed Pressman 2

Writers 137: Neil Simon 3, Alan Ball 2, Bo Goldman 2, Brian Helgeland 2, David Hare 2

Cast 323: Jack Nicholson 6, Ralph Fiennes 4, Frances Fisher 3, Leonardo DiCaprio 3, Tommy Lee Jones 3

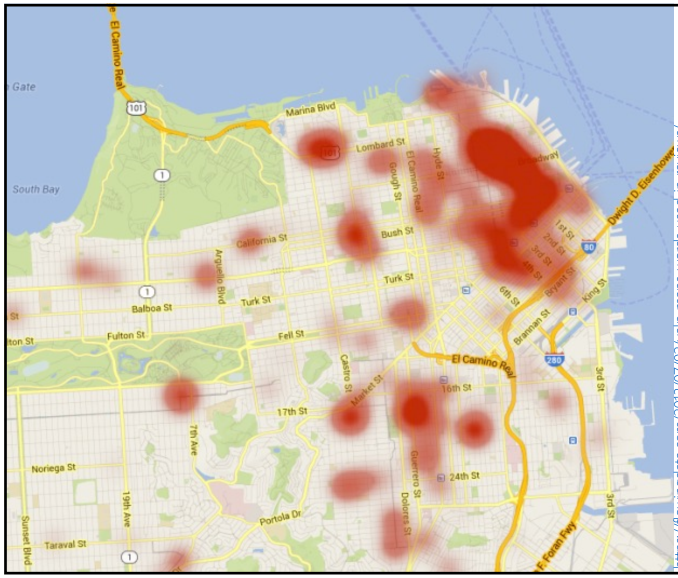
Crew 203: Bob Richardson 3, John Kander 3, Bruno Rubeo 2, Carlo Di Palma 2, Fred Ebb 2

Agent 115: Ed Limato 4, Kevin Huvane 4, Sam Cohn 4, Hylda Queally 3, Josh Lieberman 3

Publicist 45: Lisa Kasteler 4, Stephen Huvane 2, Troy Nankin 2

https://www.rebeccarolfe.com/projects/thanktheacademy/sections/network.html

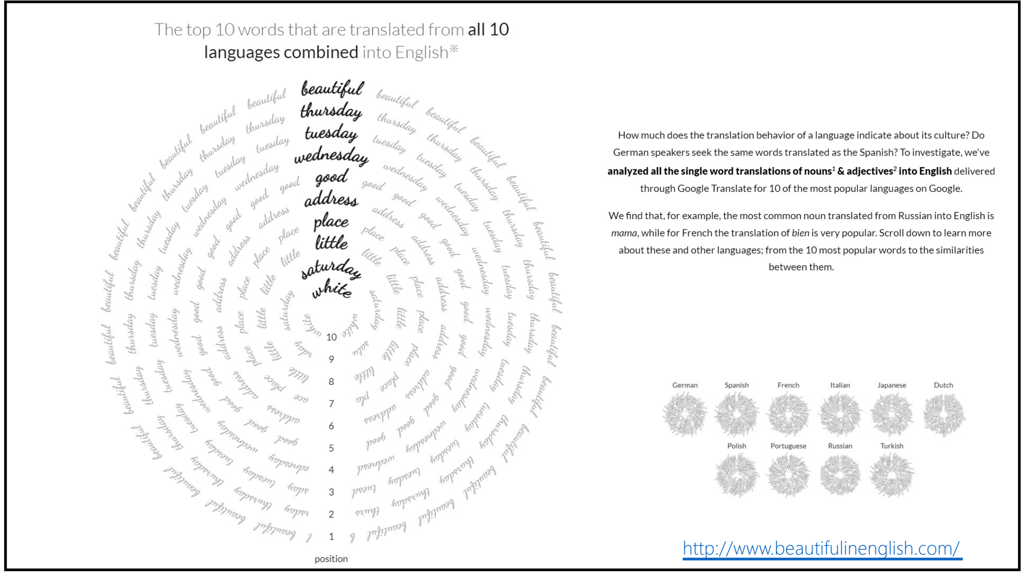
8



https://flowingdata.com/2013/07/02/yelp-maps-words-used-in-reviews/

where *"pasta"* is often used in San Francisco reviews.

9



The top 10 words that are translated from **all 10 languages combined** into English*

How much does the translation behavior of a language indicate about its culture? Do German speakers seek the same words translated as the Spanish? To investigate, we've **analyzed all the single word translations of nouns[1] & adjectives[2] into English** delivered through Google Translate for 10 of the most popular languages on Google.

We find that, for example, the most common noun translated from Russian into English is *mama*, while for French the translation of *bien* is very popular. Scroll down to learn more about these and other languages; from the 10 most popular words to the similarities between them.

http://www.beautifulinenglish.com/

12

TODAY

Learning Objectives

1. Considering text as data
2. Identifying descriptive words/keyphrases
3. Visualizing text in context
4. Searching across and comparing documents

13

TEXT AS DATA

14

TEXT AS NOMINAL DATA

High dimensional (many words, e.g. 10,000+ common words)

More than equality tests

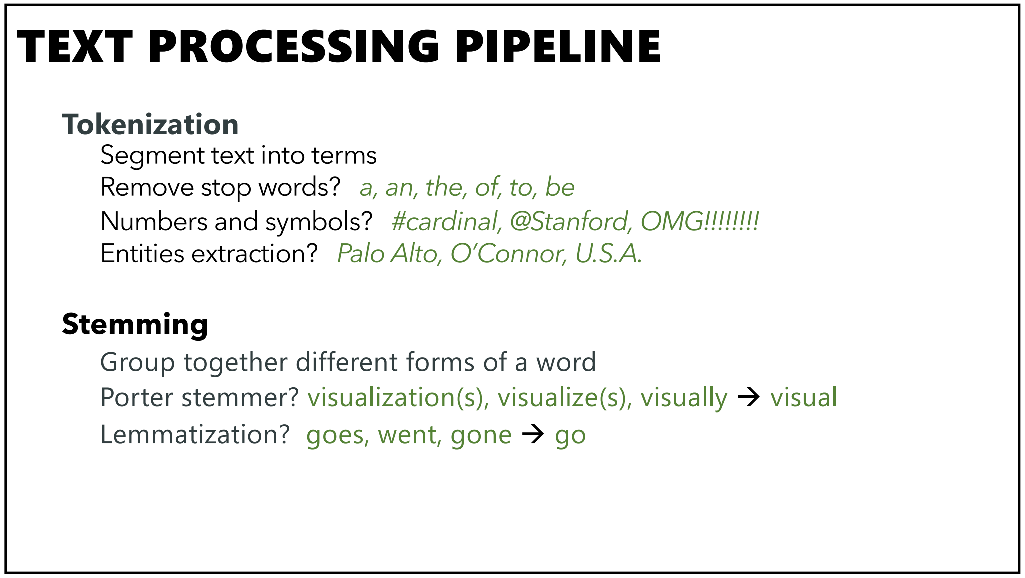
- Correlations: *Hong Kong, San Francisco, Bay Area*
- Order: *April, February, January, June, March, May*
- Membership: *Tennis, Running, Swimming, Hiking, Piano*
- Hierarchy, antonyms & synonyms, entities, …

Words have meanings and relations

TEXT PROCESSING PIPELINE

Tokenization

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



TEXT PROCESSING PIPELINE

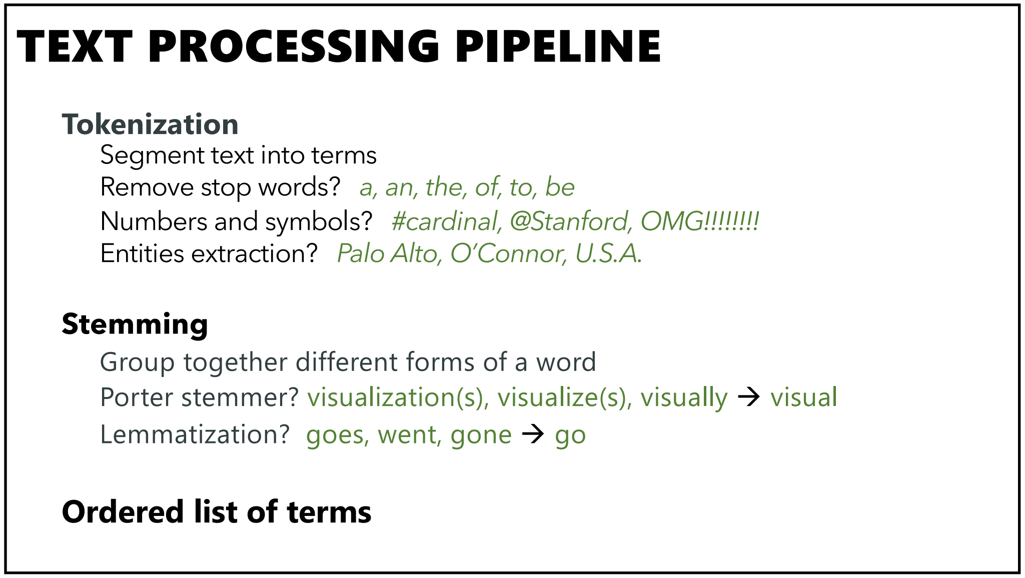
Tokenization

Segment text into terms
Remove stop words? *a, an, the, of, to, be*
Numbers and symbols? *#cardinal, @Stanford, OMG!!!!!!!!*
Entities extraction? *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

17



TEXT PROCESSING PIPELINE

Tokenization

Segment text into terms
Remove stop words? *a, an, the, of, to, be*
Numbers and symbols? *#cardinal, @Stanford, OMG!!!!!!!!*
Entities extraction? *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

18

IDENTIFYING DESCRIPTIVE WORDS/KEYPHRASES

20

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

21

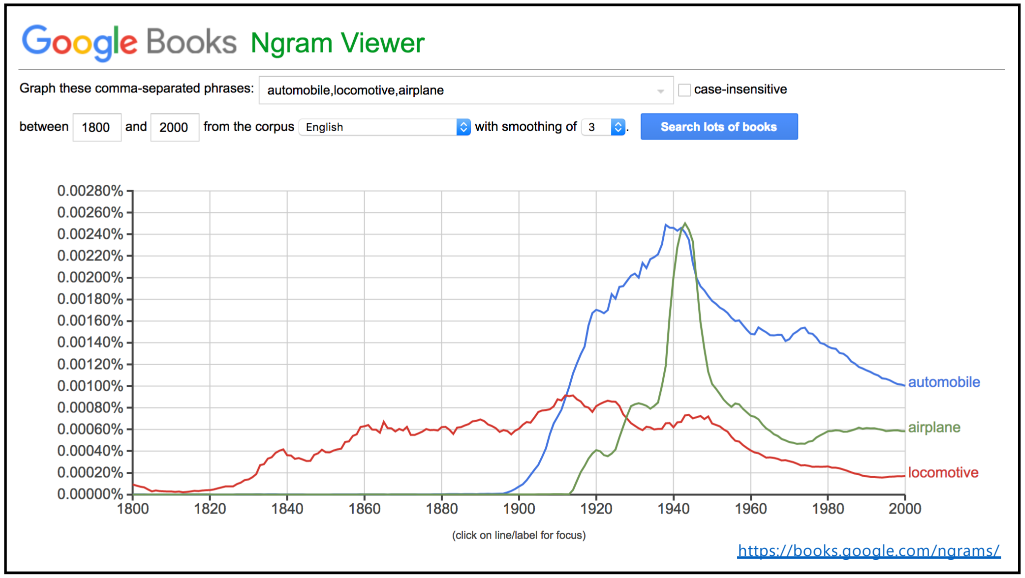
DOCUMENT X TERM MATRIX

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

22



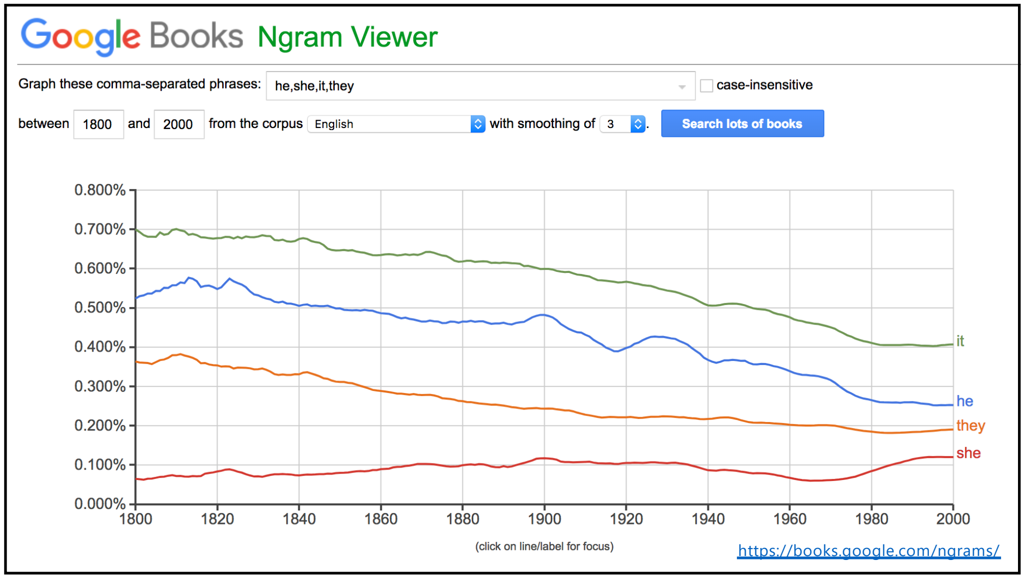
Google Books Ngram Viewer

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

https://books.google.com/ngrams/

24



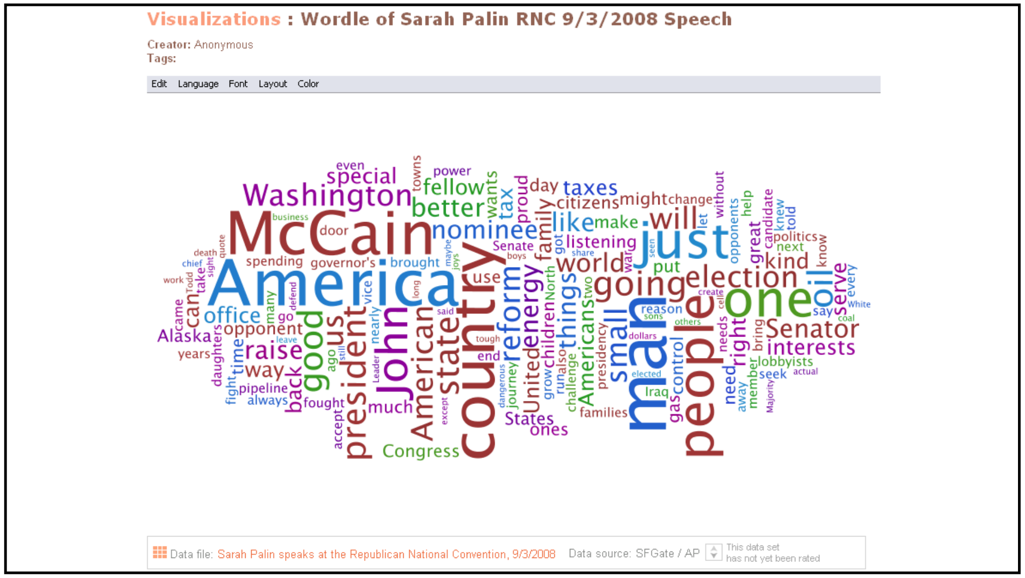
Google Books Ngram Viewer

Graph these comma-separated phrases: he,she,it,they ☐ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

https://books.google.com/ngrams/

25



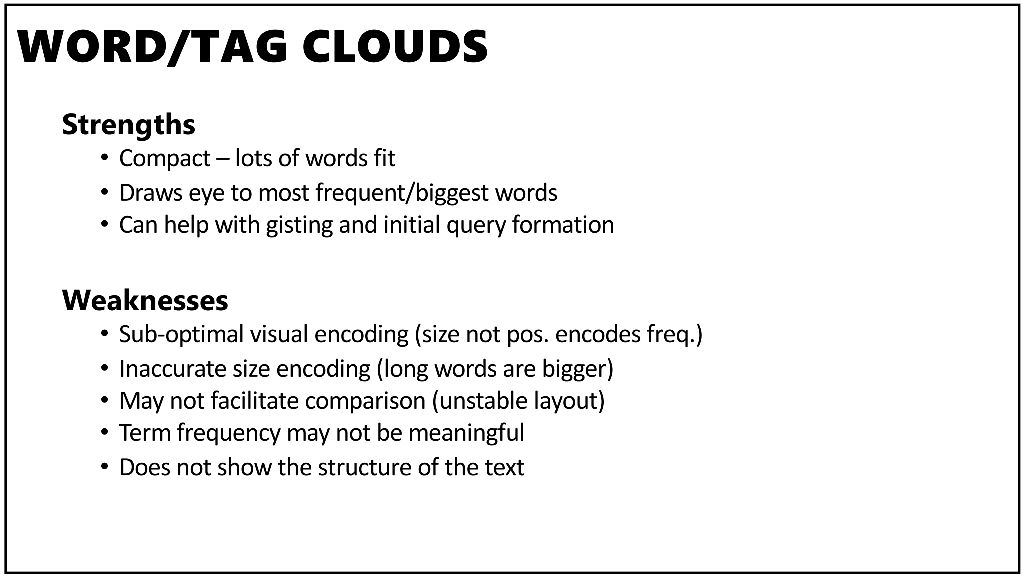
Visualizations : Wordle of Sarah Palin RNC 9/3/2008 Speech

Creator: Anonymous

Tags:

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

27



WORD/TAG CLOUDS

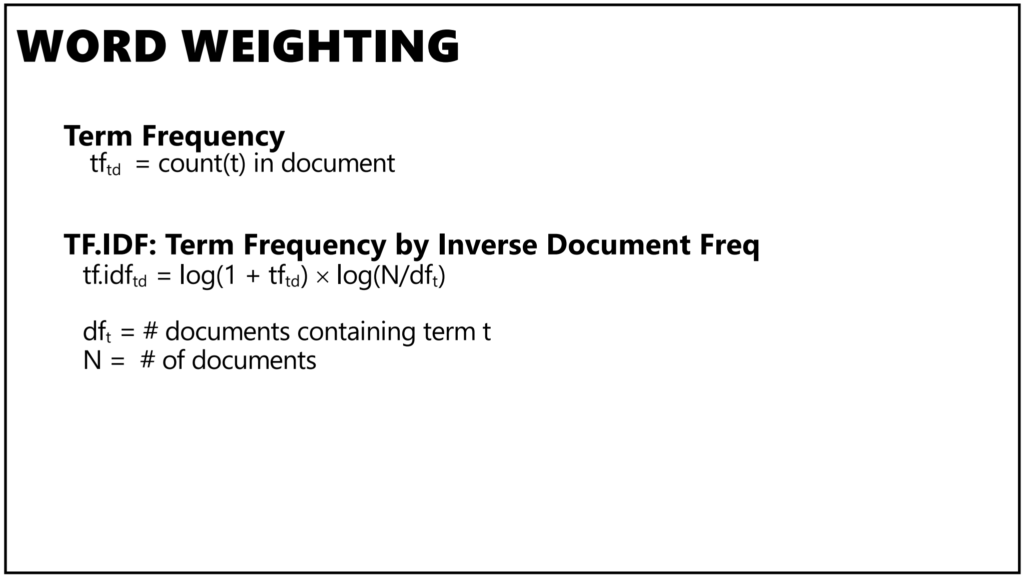
Strengths

- Compact – lots of words fit
- Draws eye to most frequent/biggest words
- 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

28



WORD WEIGHTING

Term Frequency

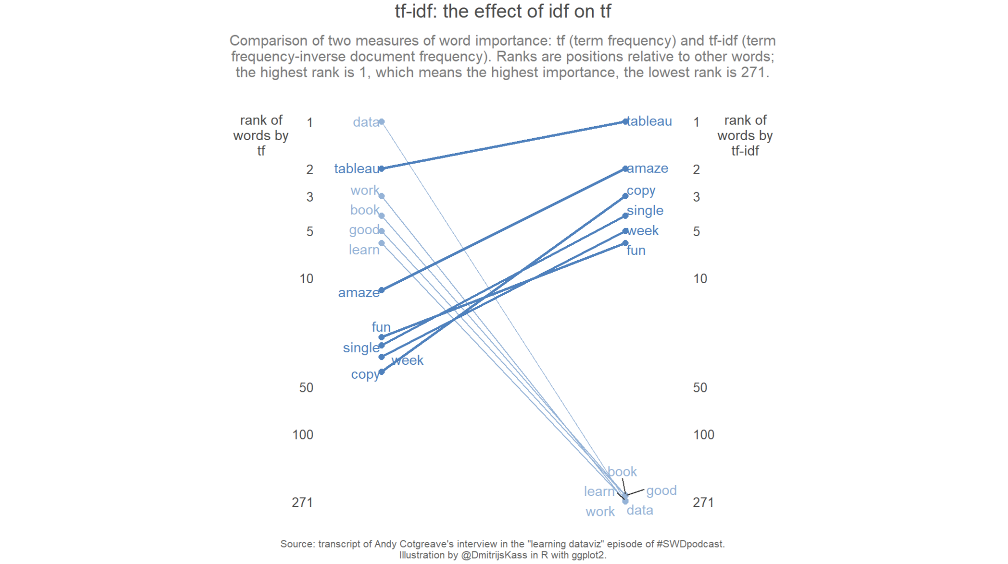
tf_{td} = count(t) in document

TF.IDF: Term Frequency by Inverse Document Freq

$\mathrm{tf.idf}_{td} = \log(1 + \mathrm{tf}_{td}) \times \log(N/\mathrm{df}_t)$

df_t = # documents containing term t
N = # of documents

29



tf-idf: the effect of idf on tf

Comparison of two measures of word importance: tf (term frequency) and tf-idf (term frequency-inverse document frequency). Ranks are positions relative to other words; the highest rank is 1, which means the highest importance, the lowest rank is 271.

Source: transcript of Andy Cotgreave's interview in the "learning dataviz" episode of #SWDpodcast.
Illustration by @DmitrijsKass in R with ggplot2.

30

LIMITATIONS OF FREQUENCY STATISTICS

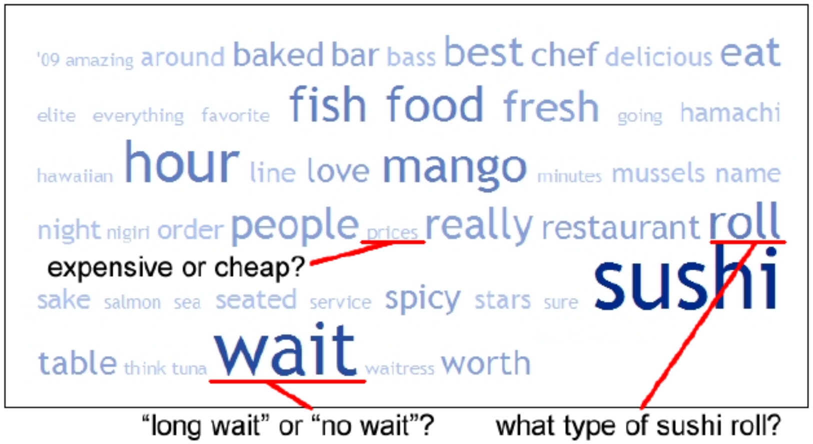
Typically focus on unigrams (single terms)

Often favors frequent (TF) or rare (IDF) terms
Not clear that these provide best description of documents

“Bag of words” ignores additional info
- Grammar / part-of-speech
- Position within document
- Recognizable entities

31

YELP REVIEW SPOTLIGHT


'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
sake salmon sea seated service spicy stars sure sushi
table think tuna wait waitress worth

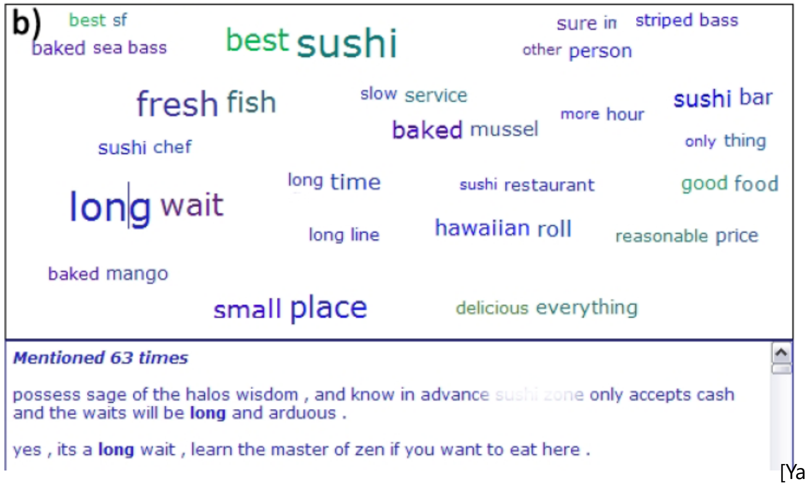
expensive or cheap?

“long wait” or “no wait”? what type of sushi roll?

[Yatani 2011]

32

YELP REVIEW SPOTLIGHT



Mentioned 63 times

possess sage of the halos wisdom , and know in advance sushi zone only accepts cash and the waits will be **long** and arduous .

yes , its a **long** wait , learn the master of zen if you want to eat here .

[Yatani 2011]

33

TIPS FOR 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

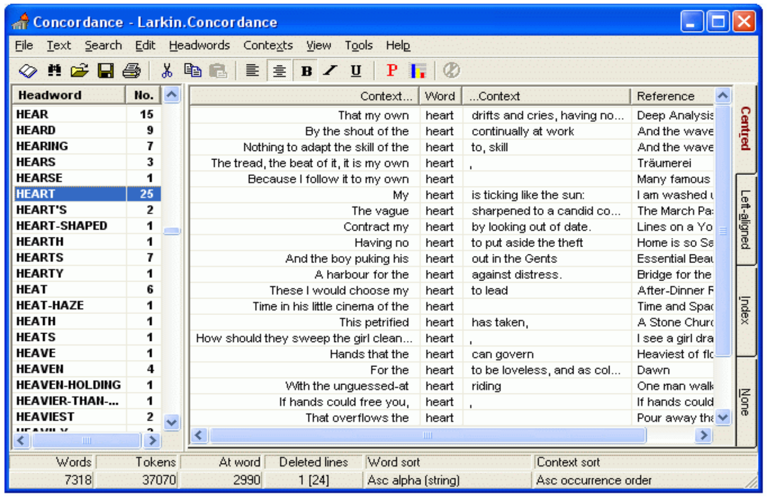
- Generate longer, more meaningful phrases
- Adjective-noun word pairs for reviews

34

VISUALIZING WORDS/KEYPHRASES CONTEXT

35

CONCORDANCE



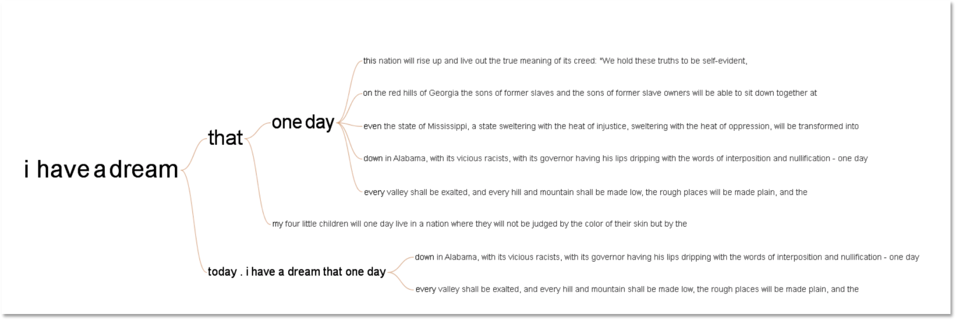
Headword	No.
HEAR	15
HEARD	9
HEARING	7
HEARS	3
HEARSE	1
HEART	25
HEART'S	2
HEART-SHAPED	1
HEARTH	1
HEARTS	7
HEARTY	1
HEAT	6
HEAT-HAZE	1
HEATH	1
HEATS	1
HEAVE	1
HEAVEN	4
HEAVEN-HOLDING	1
HEAVIER-THAN-...	1
HEAVIEST	2

Context...	Word	...Context	Reference
That my own	heart	drifts and cries, having no...	Deep Analysis
By the shout of the	heart	continually at work	And the wave
Nothing to adapt the skill of the	heart	to, skill	And the wave
The tread, the beat of it, it is my own	heart	,	Träumerei
Because I follow it to my own	heart		Many famous
My	heart	is ticking like the sun:	I am washed
The vague	heart	sharpened to a candid co...	The March Pa
Contract my	heart	by looking out of date.	Lines on a Yo
Having no	heart	to put aside the theft	Home is so Sa
And the boy puking his	heart	out in the Gents	Essential Beau
A harbour for the	heart	against distress.	Bridge for the
These I would choose my	heart	to lead	After-Dinner R
Time in his little cinema of the	heart		Time and Spac
This petrified	heart	has taken,	A Stone Churc
How should they sweep the girl clean...	heart	,	I see a girl dra
Hands that the	heart	can govern	Heaviest of flo
For the	heart	to be loveless, and as col...	Dawn
With the unguessed-at	heart	riding	One man walk
If hands could free you,	heart	,	If hands could
That overflows the	heart		Pour away tha

Words	Tokens	At word	Deleted lines	Word sort	Context sort
7318	37070	2990	1 [24]	Asc alpha (string)	Asc occurrence order

www.concordancesoftware.co.uk

WORD TREE



i have a dream

that

one day

this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident,

on the red hills of Georgia the sons of former slaves and the sons of former slave owners will be able to sit down together at

even the state of Mississippi, a state sweltering with the heat of injustice, sweltering with the heat of oppression, will be transformed into

down in Alabama, with its vicious racists, with its governor having his lips dripping with the words of interposition and nullification - one day

every valley shall be exalted, and every hill and mountain shall be made low, the rough places will be made plain, and the

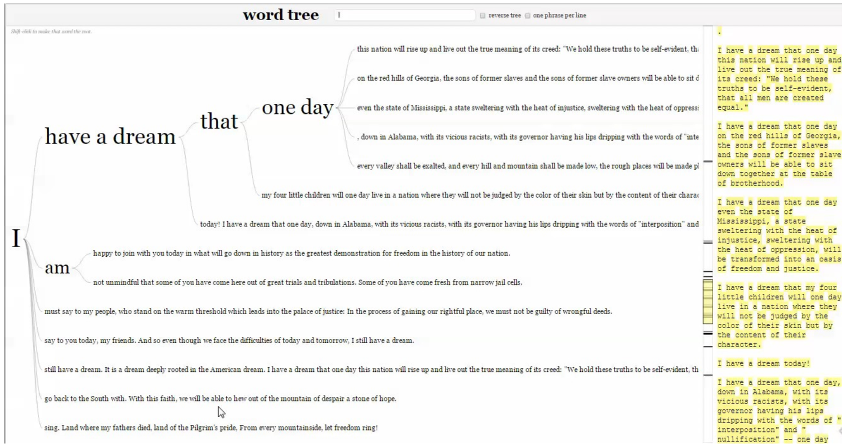
my four little children will one day live in a nation where they will not be judged by the color of their skin but by the

today . i have a dream that one day

down in Alabama, with its vicious racists, with its governor having his lips dripping with the words of interposition and nullification - one day

every valley shall be exalted, and every hill and mountain shall be made low, the rough places will be made plain, and the

The Word Tree, An Interactive Visual Concordance [pdf]
Martin Wattenberg & Fernanda Viégas, InfoVis 2008.



this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident, th

on the red hills of Georgia, the sons of former slaves and the sons of former slave owners will be able to sit d

even the state of Mississippi, a state sweltering with the heat of injustice, sweltering with the heat of oppress

, down in Alabama, with its vicious racists, with its governor having his lips dripping with the words of "inte

every valley shall be exalted, and every hill and mountain shall be made low, the rough places will be made pl

my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their charac

today! I have a dream that one day, down in Alabama, with its vicious racists, with its governor having his lips dripping with the words of "interposition" and

happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation.

not unmindful that some of you have come here out of great trials and tribulations. Some of you have come fresh from narrow jail cells.

must say to my people, who stand on the warm threshold which leads into the palace of justice: In the process of gaining our rightful place, we must not be guilty of wrongful deeds.

say to you today, my friends. And so even though we face the difficulties of today and tomorrow, I still have a dream.

still have a dream. It is a dream deeply rooted in the American dream. I have a dream that one day this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident, th

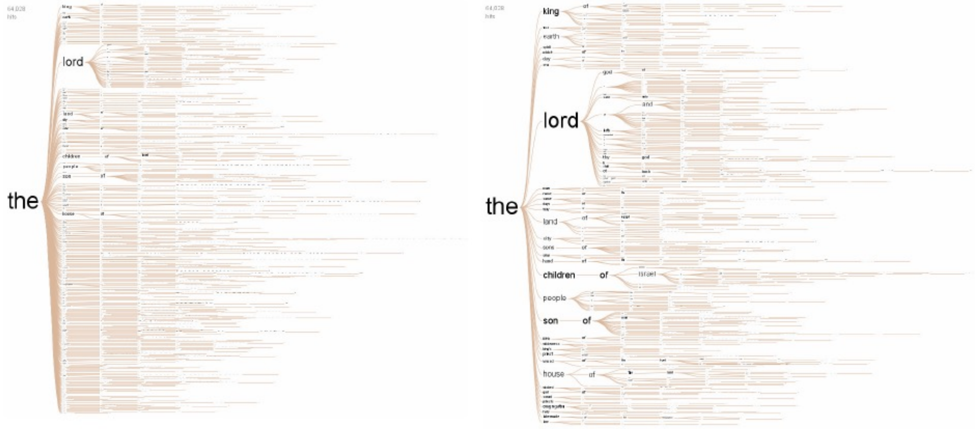
go back to the South with. With this faith, we will be able to hew out of the mountain of despair a stone of hope.

sing. Land where my fathers died, land of the Pilgrim's pride, From every mountainside, let freedom ring!

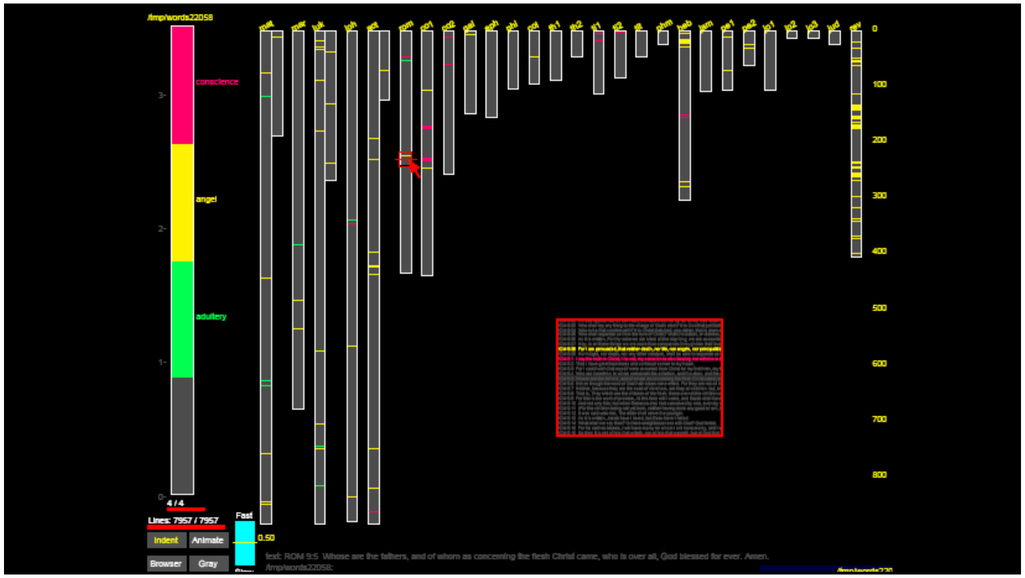
```
.

I have a dream that one day
this nation will rise up and
live out the true meaning of
its creed: "We hold these
truths to be self-evident,
that all men are created
equal."

I have a dream that one day
on the red hills of Georgia,
the sons of former slaves
and the sons of former slave
owners will be able to sit
down together at the table
of brotherhood.

I have a dream that one day
even the state of
Mississippi, a state
sweltering with the heat of
injustice, sweltering with
the heat of oppression, will
be transformed into an oasis
of freedom and justice.

I have a dream that my four
little children will one day
live in a nation where they
will not be judged by the
color of their skin but by
the content of their
character.

I have a dream today!

I have a dream that one day,
down in Alabama, with its
vicious racists, with its
governor having his lips
dripping with the words of "
interposition" and "
nullification" -- one day
```

The Word Tree, An Interactive Visual Concordance [pdf]
Martin Wattenberg & Fernanda Viégas, InfoVis 2008.

FILTER INFREQUENT RUNS



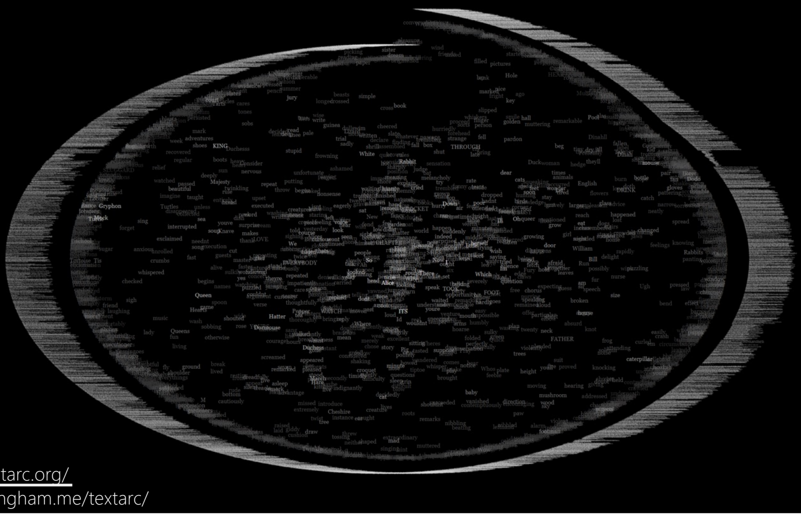
The Word Tree, An Interactive Visual Concordance [pdf]
Martin Wattenberg & Fernanda Viégas, InfoVis 2008.



text: ROM 9:5 Whose are the fathers, and of whom as concerning the flesh Christ came, who is over all, God blessed for ever. Amen.
/tmp/words22058:

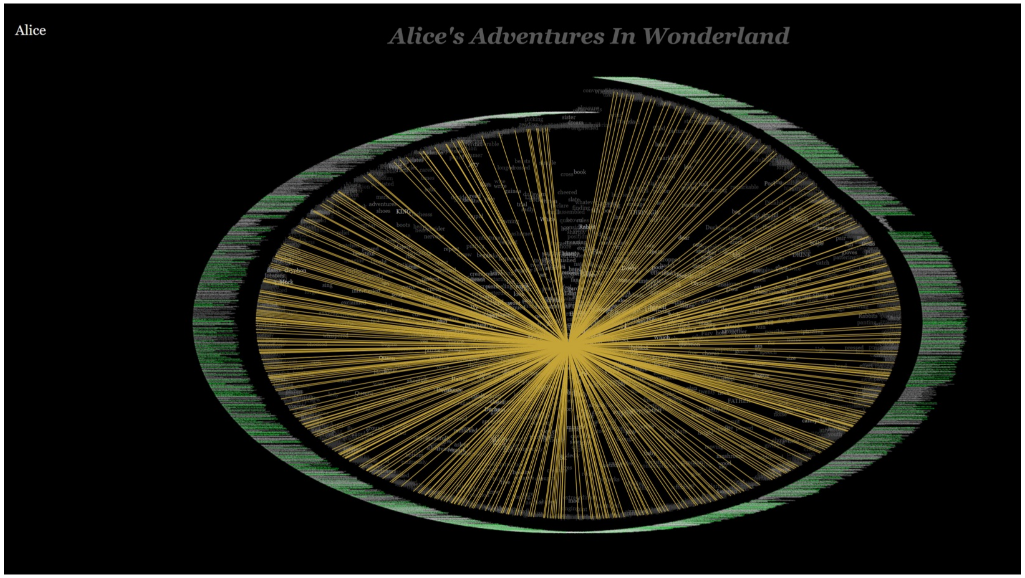
46

Alice's Adventures In Wonderland

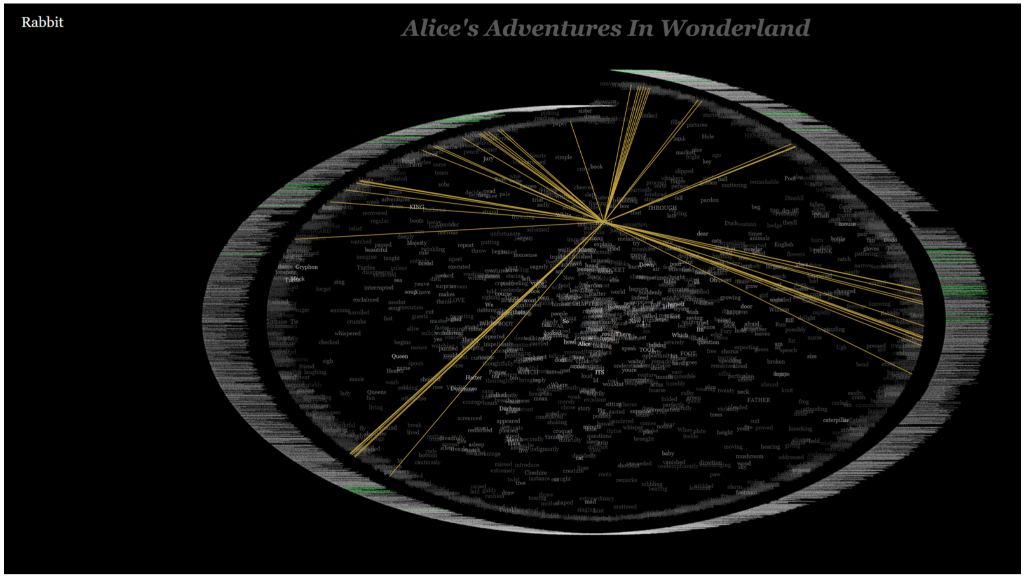


http://www.textarc.org/
http://vallandingham.me/textarc/

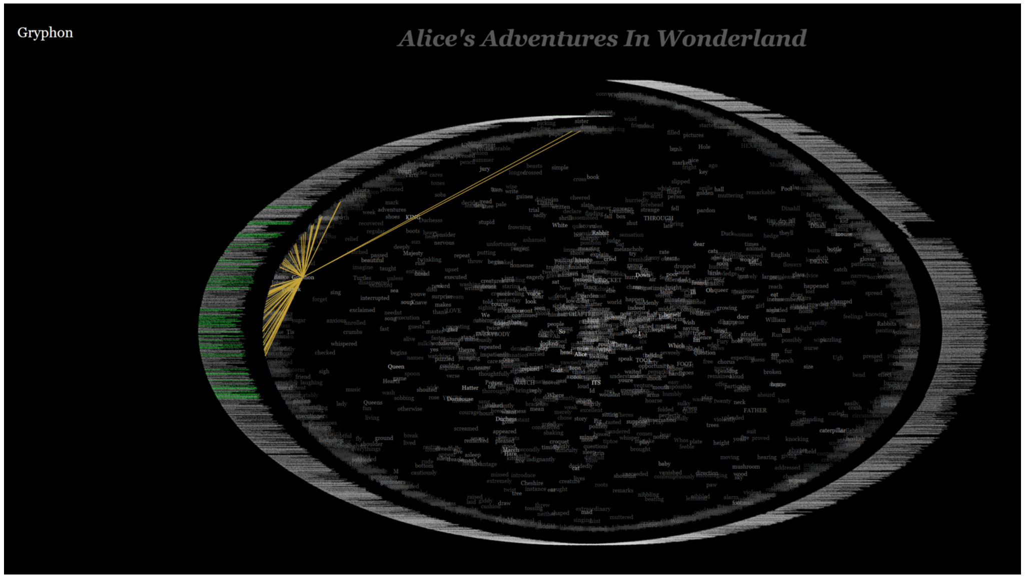
48


Alice
Alice's Adventures In Wonderland

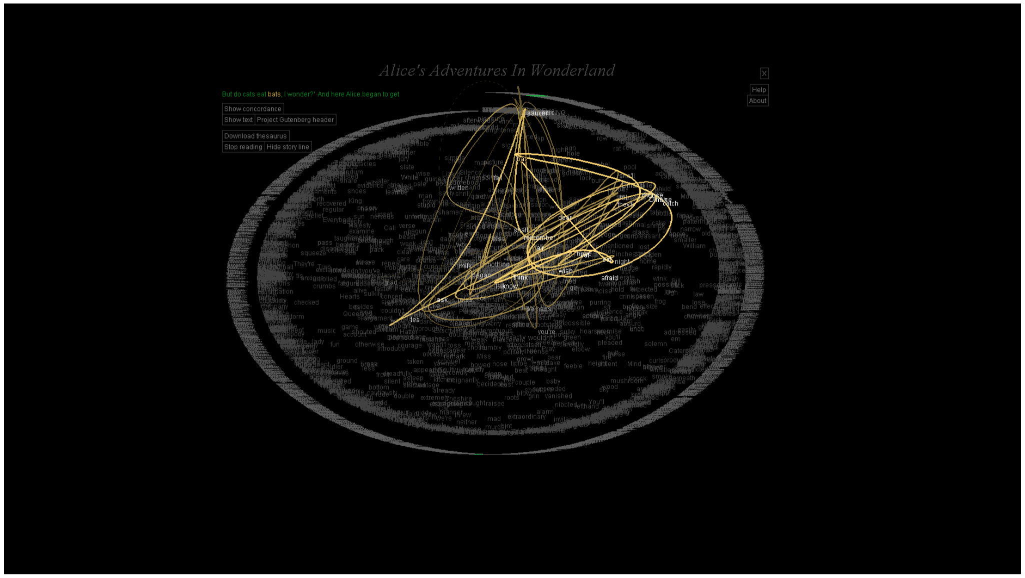
49


Rabbit
Alice's Adventures In Wonderland

50



51



But do cats eat bats, I wonder?' And here Alice began to get

52

GLIMPSES OF STRUCTURE

Concordances and TextArc show local, repeated structure

But what about other types of patterns?

For example

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

PHRASE NETS

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

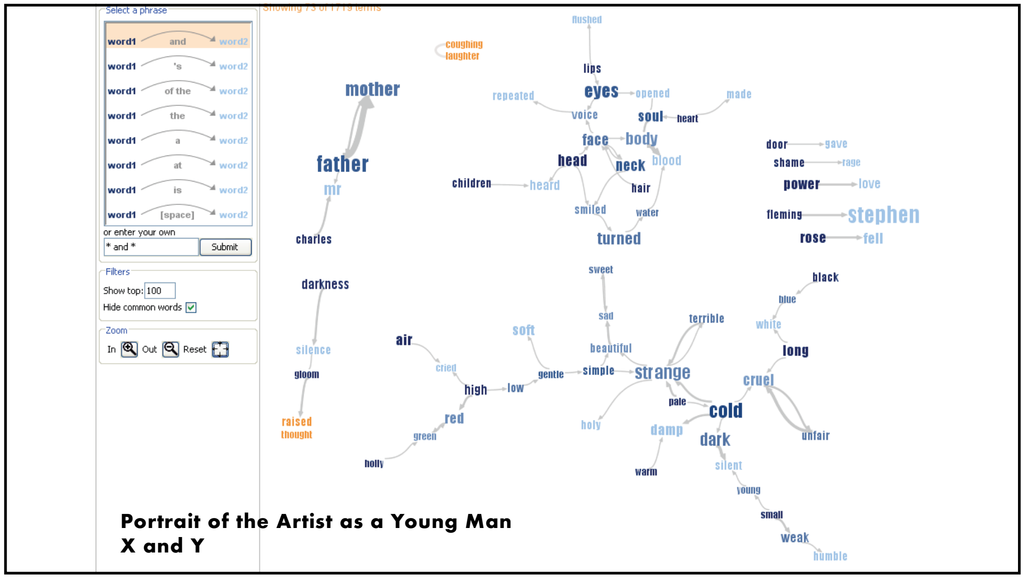
Visualize extracted patterns in a node-link view

Occurrences → Node size

Pattern position → Edge direction

Darker color → higher ratio of out-edges to in-edges

Mapping Text with Phrase Nets
Frank Van Ham, Martin Wattenberg & Fernanda Viégas, InfoVis 2009.



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

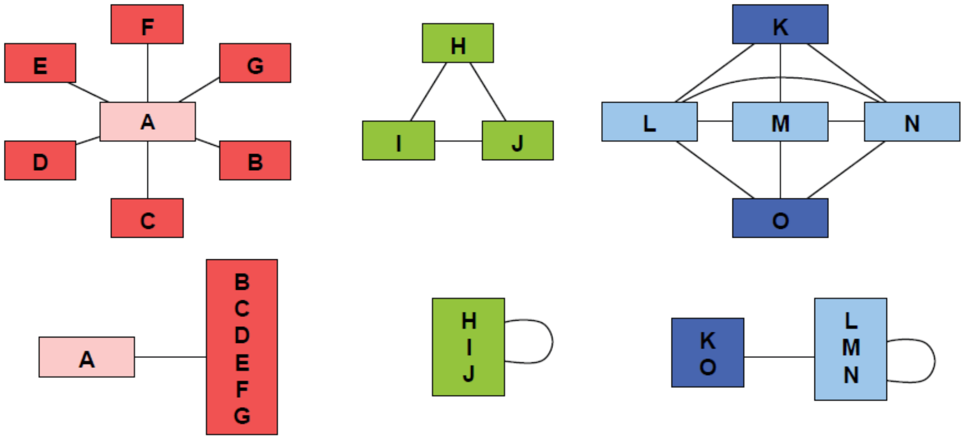
In Out Reset

Showing 73 of 1719 terms

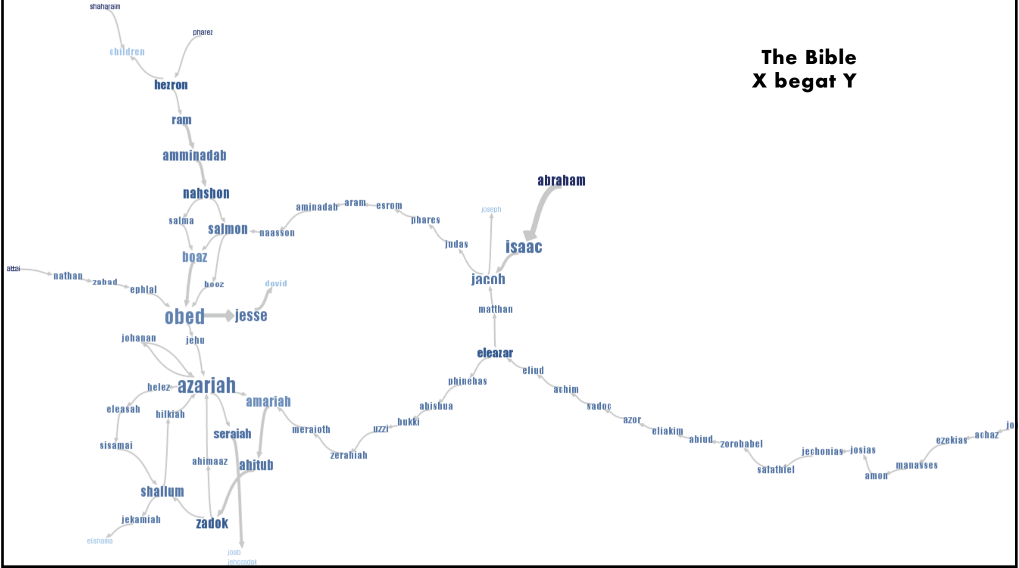
Portrait of the Artist as a Young Man
X and Y

56

EDGE COMPRESSION/NODE GROUPING

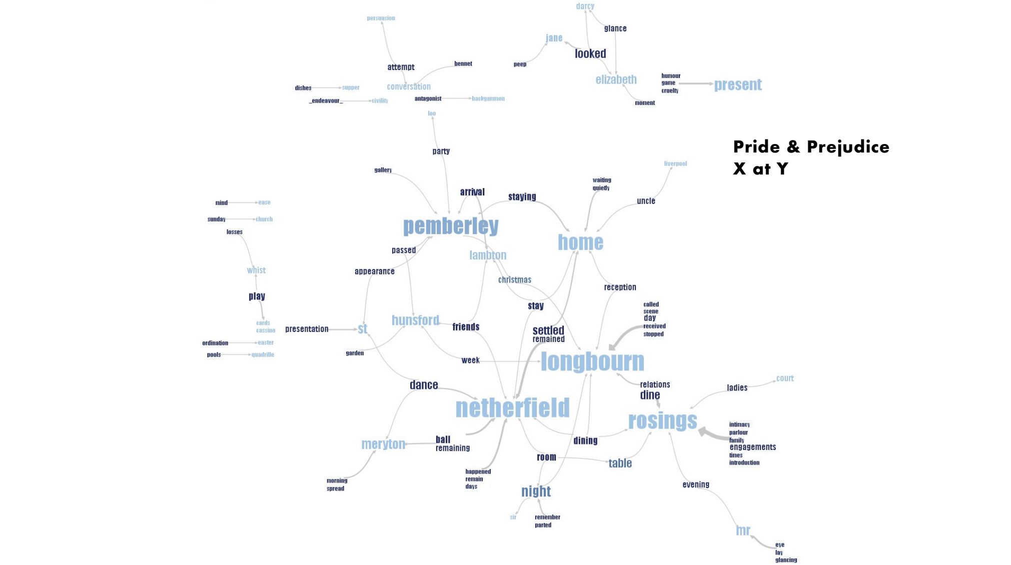


57



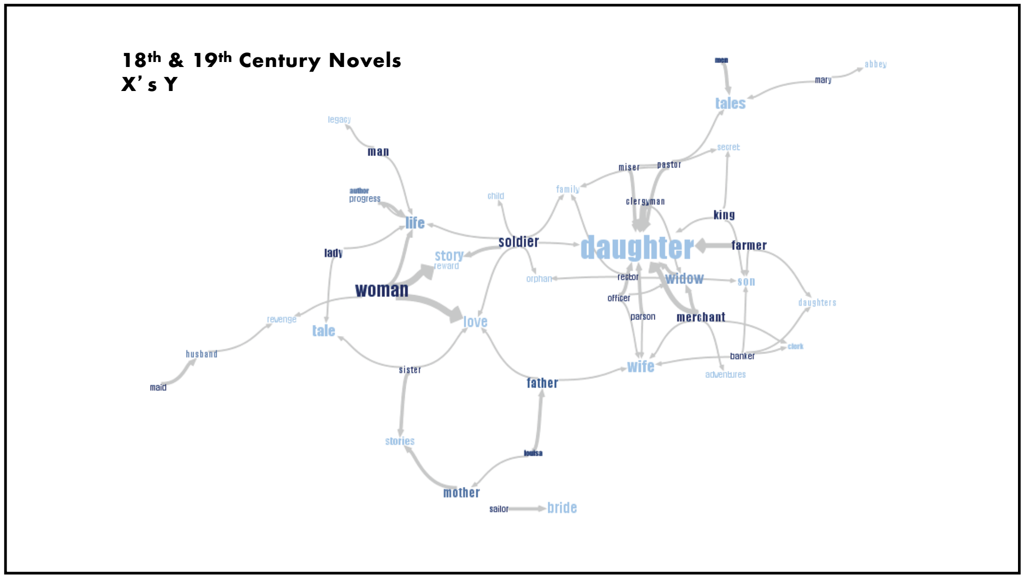
The Bible
X begat Y

58



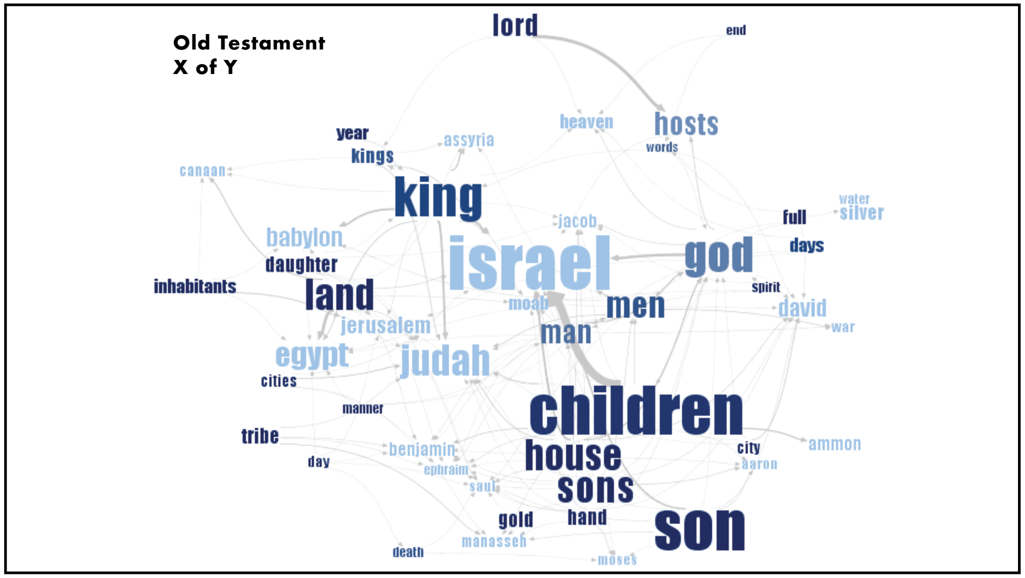
Pride & Prejudice
X at Y

59



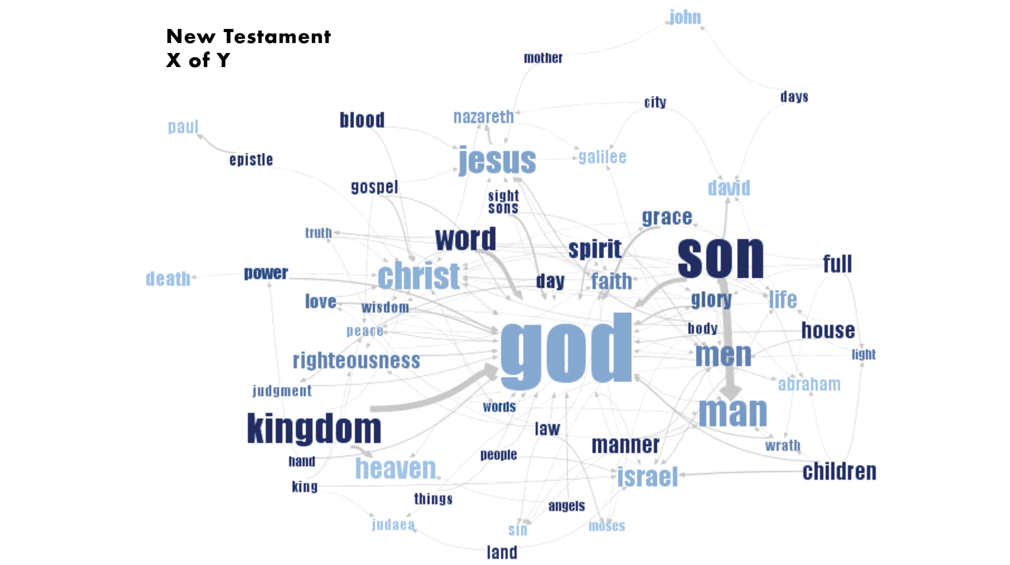
18th & 19th Century Novels
X' s Y

62



Old Testament
X of Y

63

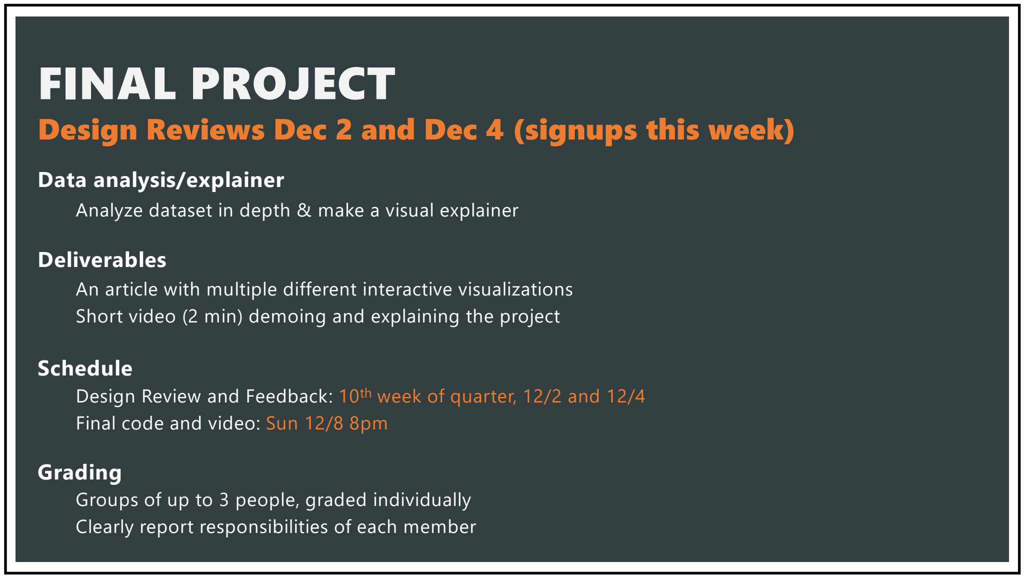


New Testament
X of Y

64

ANNOUNCEMENTS

65



FINAL PROJECT

Design Reviews Dec 2 and Dec 4 (signups this week)

Data analysis/explainer

- Analyze dataset in depth & make a visual explainer

Deliverables

- An article with multiple different interactive visualizations
- Short video (2 min) demoing and explaining the project

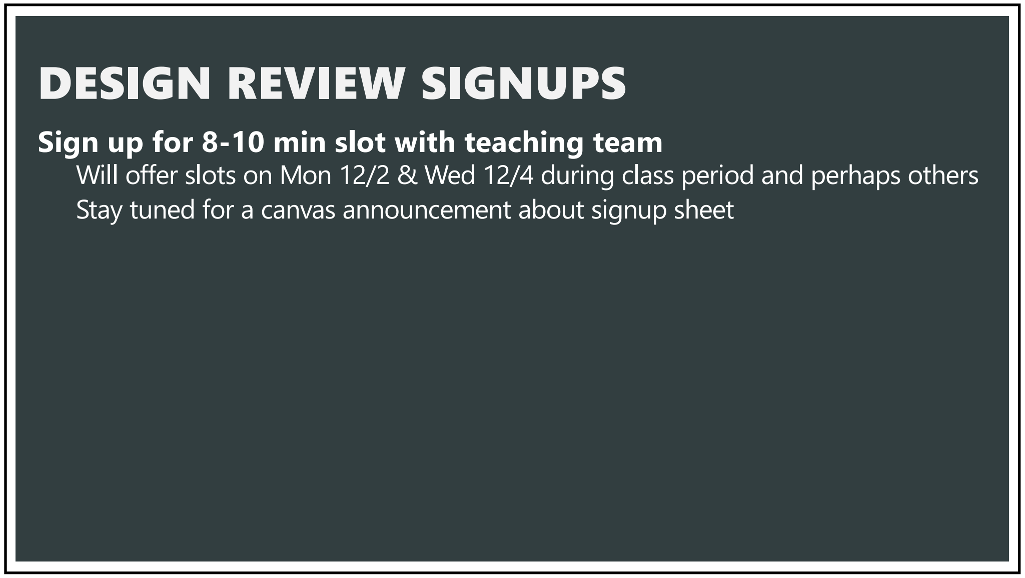
Schedule

- Design Review and Feedback: 10th week of quarter, 12/2 and 12/4
- Final code and video: Sun 12/8 8pm

Grading

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

66



DESIGN REVIEW SIGNUPS

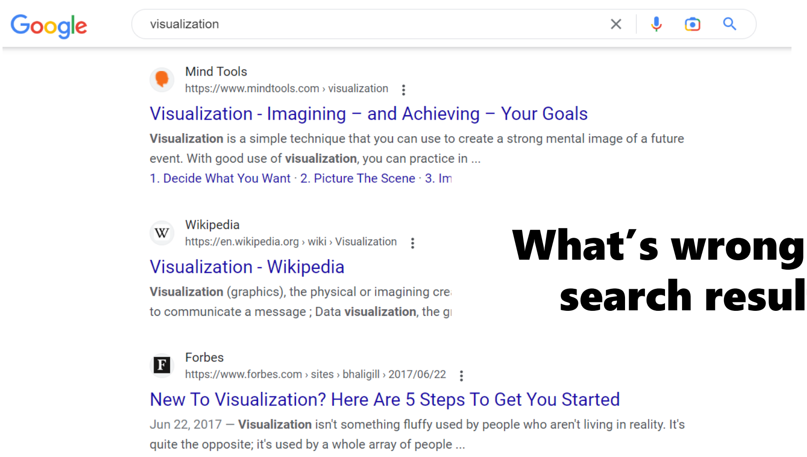
Sign up for 8-10 min slot with teaching team

- Will offer slots on Mon 12/2 & Wed 12/4 during class period and perhaps others
- Stay tuned for a canvas announcement about signup sheet

67

SEARCHING ACROSS DOCUMENT(S)

68



Mind Tools
https://www.mindtools.com › visualization

Visualization - Imagining – and Achieving – Your Goals

Visualization is a simple technique that you can use to create a strong mental image of a future event. With good use of **visualization**, you can practice in ...

1. Decide What You Want · 2. Picture The Scene · 3. Im

Wikipedia
https://en.wikipedia.org › wiki › Visualization

Visualization - Wikipedia

Visualization (graphics), the physical or imagining cre
to communicate a message ; Data **visualization**, the g

Forbes
https://www.forbes.com › sites › bhaligill › 2017/06/22

New To Visualization? Here Are 5 Steps To Get You Started

Jun 22, 2017 — **Visualization** isn't something fluffy used by people who aren't living in reality. It's quite the opposite; it's used by a whole array of people ...

What's wrong with search results?

69

PROBLEMS WITH SEARCH RESULTS

Query responses don't tell you:

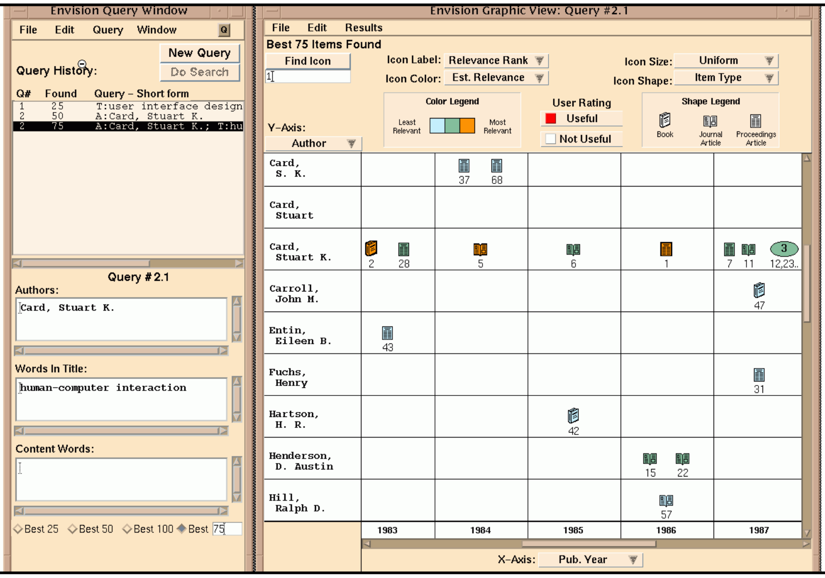
- How strong the match
- How frequent each term
- How term is distributed
- Overlap between terms
- Length of document

Ranking is opaque

Inability to compare results

Marti A. Hearst

70



Query History:

Q#	Found	Query – Short form
1	25	T:user interface design
2	50	A:Card, Stuart K.
2	75	A:Card, Stuart K.; T:hu

Query # 2.1

Authors: Card, Stuart K.

Words In Title: human-computer interaction

Content Words:

◇ Best 25 ◇ Best 50 ◇ Best 100 ◆ Best 75

Best 75 Items Found

1

Icon Label: Relevance Rank

Icon Color: Est. Relevance

Icon Size: Uniform

Icon Shape: Item Type

User Rating: 🟥 Useful □ Not Useful

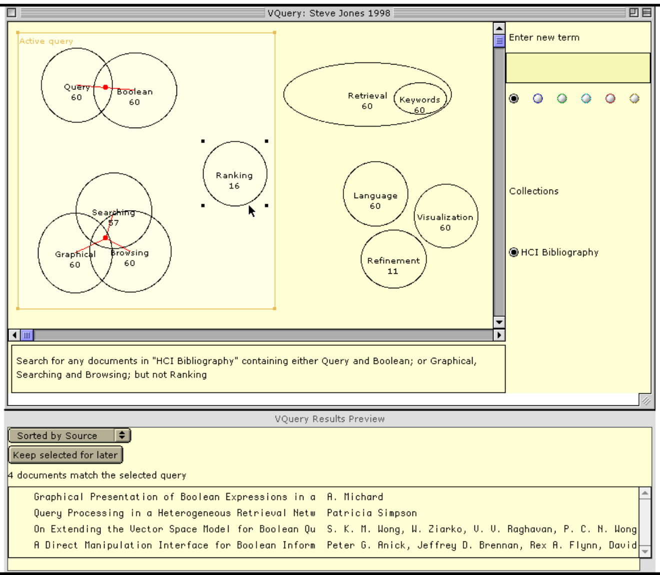
Y-Axis: Author

	1983	1984	1985	1986	1987
Card, S. K.		37 68			
Card, Stuart					
Card, Stuart K.	2 28	5	6	1	7 11 3 12,23..
Carroll, John M.					47
Entin, Eileen B.	43				
Fuchs, Henry					31
Hartson, H. R.			42		
Henderson, D. Austin				15 22	
Hill, Ralph D.				57	

X-Axis: Pub. Year

Fox et. al., Users, user interfaces, and objects: Envision, a digital library. *JASIS*, 44(8):480-491, 1993.

71



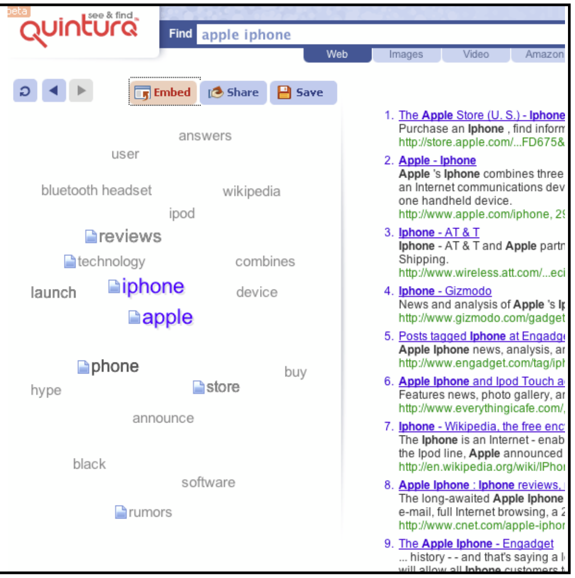
Search for any documents in "HCI Bibliography" containing either Query and Boolean; or Graphical, Searching and Browsing; but not Ranking

4 documents match the selected query

```
Graphical Presentation of Boolean Expressions in a  A. Michard
Query Processing in a Heterogeneous Retrieval Netw  Patricia Simpson
On Extending the Vector Space Model for Boolean Qu  S. K. M. Wong, W. Ziarko, V. V. Raghavan, P. C. N. Wong
A Direct Manipulation Interface for Boolean Inform  Peter G. Anick, Jeffrey D. Brennan, Rex A. Flynn, David
```

VQuery
Jones & McInnes, UIST'98

72

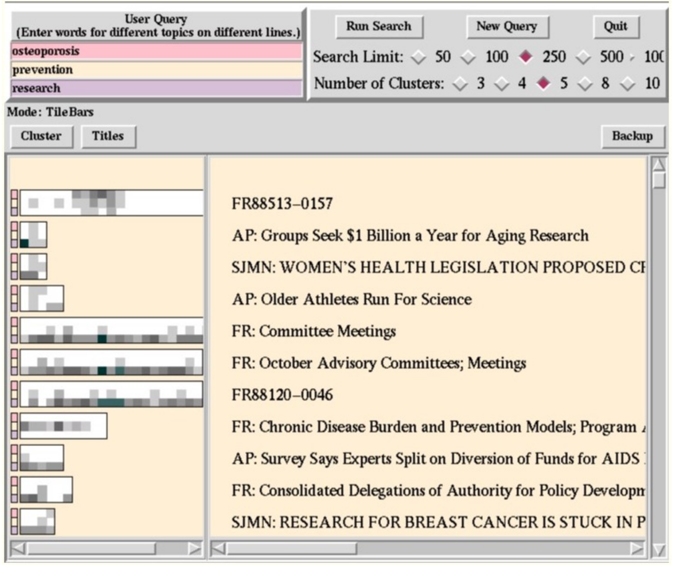


1. The Apple Store (U. S.) - Iphone
Purchase an Iphone , find inform
http://store.apple.com/...FD675&
2. Apple - Iphone
Apple 's Iphone combines three
an Internet communications dev
one handheld device.
http://www.apple.com/iphone, 29
3. Iphone - AT & T
Iphone - AT & T and Apple partn
Shipping.
http://www.wireless.att.com/...eci
4. Iphone - Gizmodo
News and analysis of Apple 's Ip
http://www.gizmodo.com/gadget
5. Posts tagged Iphone at Engadge
Apple Iphone news, analysis, an
http://www.engadget.com/tag/iph
6. Apple Iphone and Ipod Touch a
Features news, photo gallery, an
http://www.everythingicafe.com/,
7. Iphone - Wikipedia, the free enc
The Iphone is an Internet - enab
the Ipod line, Apple announced
http://en.wikipedia.org/wiki/IPhon
8. Apple Iphone : Iphone reviews,
The long-awaited Apple Iphone
e-mail, full Internet browsing, a 2
http://www.cnet.com/apple-iphon
9. The Apple Iphone - Engadget
... history - - and that's saying a l
will allow all Iphone customers t

Query Term
Suggestion
Visualization

Quintura

73



osteoporosis
prevention
research

Search Limit: ◇ 50 ◇ 100 ◆ 250 ◇ 500 ◇ 100

Number of Clusters: ◇ 3 ◇ 4 ◆ 5 ◇ 8 ◇ 10

Mode: TileBars

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

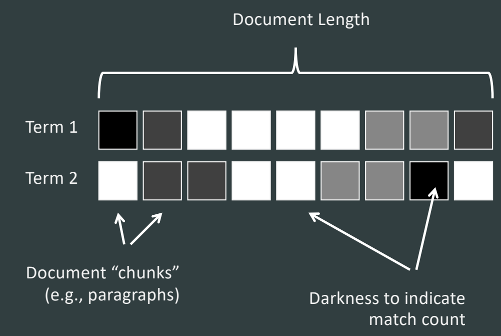
FR: Consolidated Delegations of Authority for Policy Developm

SJMN: RESEARCH FOR BREAST CANCER IS STUCK IN P

Objective: Minimize time to decide which documents to look at

TileBars, Marti A. Hearst

74



Document Length

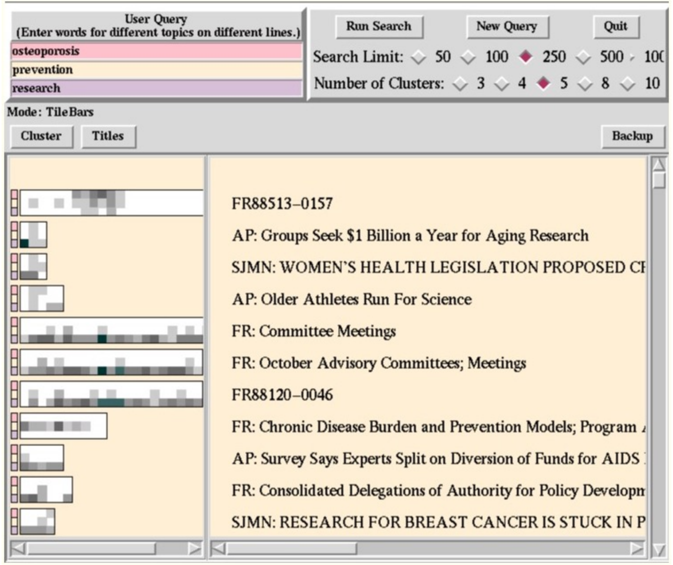
Term 1

Term 2

Document "chunks" (e.g., paragraphs)

Darkness to indicate match count

75



User Query
(Enter words for different topics on different lines.)

osteoporosis
prevention
research

Search Limit: ◇ 50 ◇ 100 ◆ 250 ◇ 500 ◇ 100

Number of Clusters: ◇ 3 ◇ 4 ◆ 5 ◇ 8 ◇ 10

Mode: TileBars

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 Developm
SJMN: RESEARCH FOR BREAST CANCER IS STUCK IN P

Objective: Minimize time to decide which documents to look at

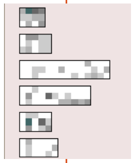
TileBars, Marti A. Hearst

76

THE POSITIVE

Simultaneous indication of:

- Relative document lengths
- Frequency of term sets in document
- Distribution of term sets with respect to the document and each other



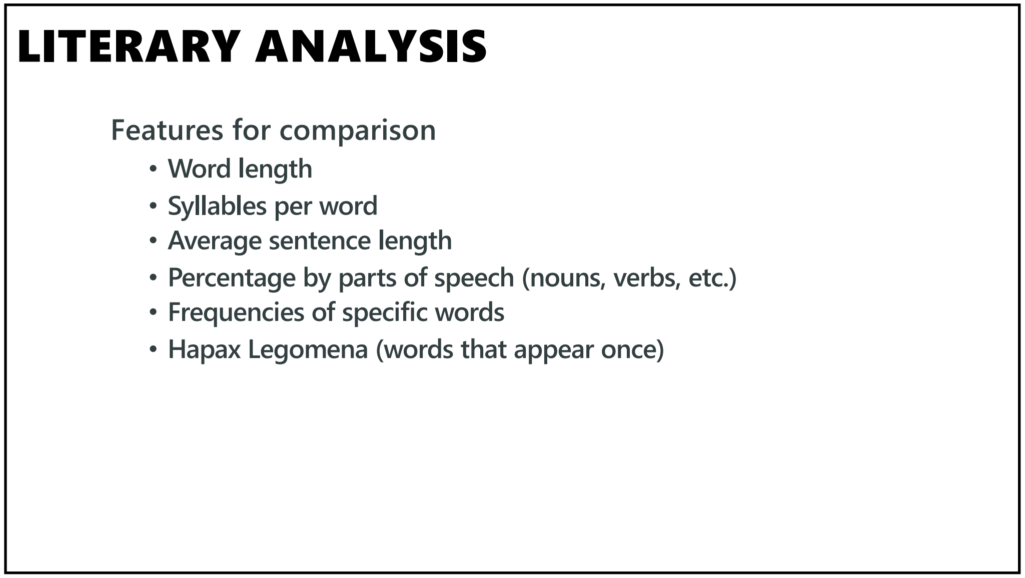
THE NEGATIVE

- Requires training to understand
- Multiple representations and doc formats *(images, page layouts, etc.)*

77

COMPARING DOCUMENTS

81

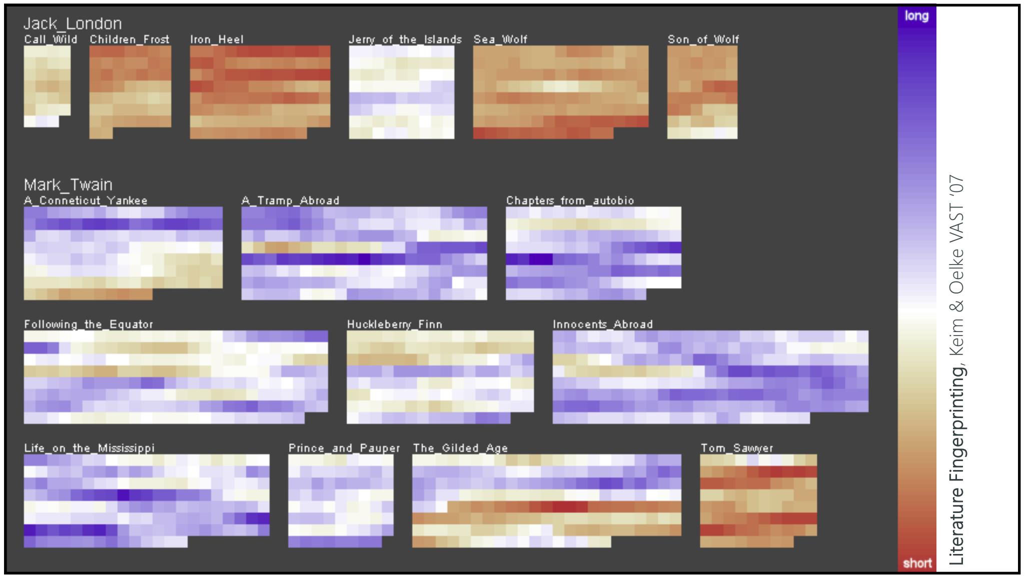


LITERARY ANALYSIS

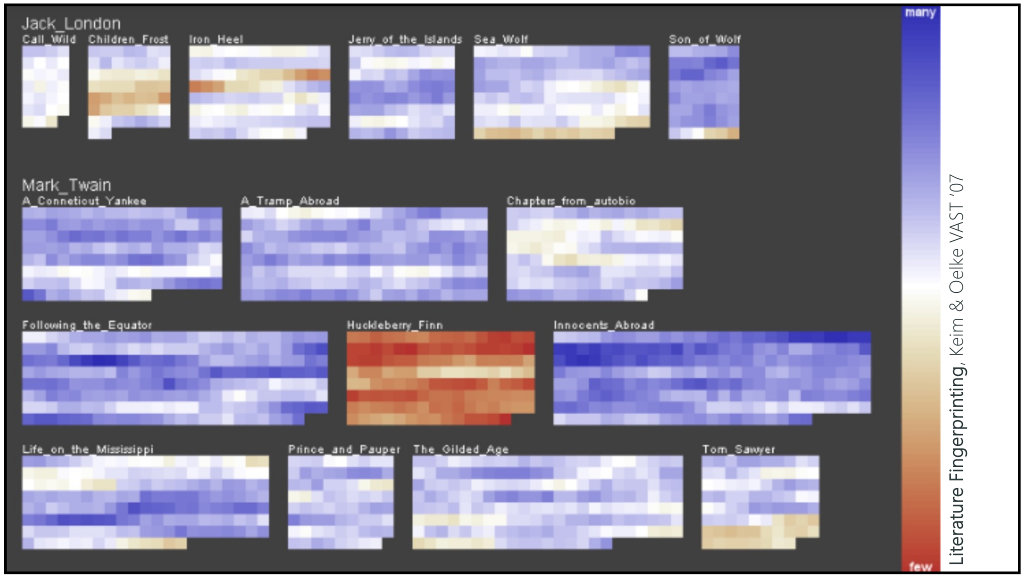
Features for comparison

- Word length
- Syllables per word
- Average sentence length
- Percentage by parts of speech (nouns, verbs, etc.)
- Frequencies of specific words
- Hapax Legomena (words that appear once)

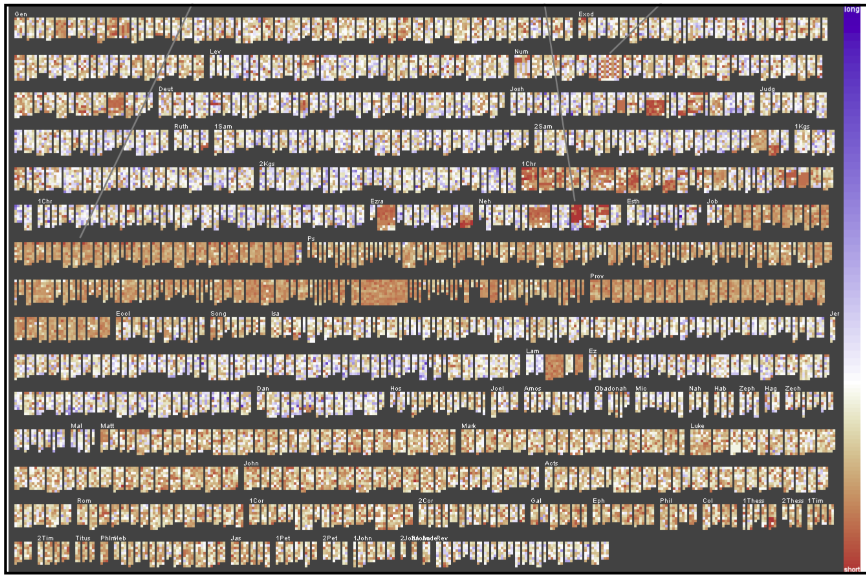
82


Jack_London
Call_Wild
Children_Frost
Iron_Heel
Jerry_of_the_Islands
Sea_Wolf
Son_of_Wolf
Mark_Twain
A_Conneticut_Yankee
A_Tramp_Abroad
Chapters_from_autobio
Following_the_Equator
Huckleberry_Finn
Innocents_Abroad
Life_on_the_Mississippi
Prince_and_Pauper
The_Gilded_Age
Tom_Sawyer
long
short
Literature Fingerprinting, Keim & Oelke VAST '07

83

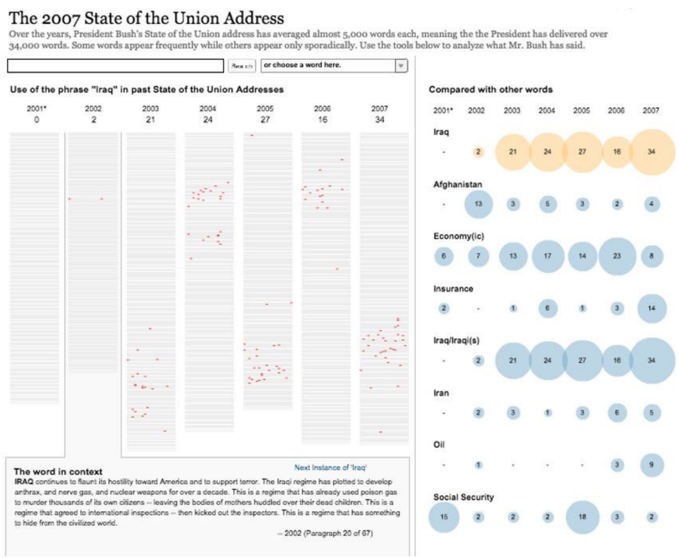

Jack_London
Call_Wild
Children_Frost
Iron_Heel
Jerry_of_the_Islands
Sea_Wolf
Son_of_Wolf
Mark_Twain
A_Conneticut_Yankee
A_Tramp_Abroad
Chapters_from_autobio
Following_the_Equator
Huckleberry_Finn
Innocents_Abroad
Life_on_the_Mississippi
Prince_and_Pauper
The_Gilded_Age
Tom_Sawyer
many
few
Literature Fingerprinting, Keim & Oelke VAST '07

84



Literature Fingerprinting, Keim & Oelke VAST '07

85



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.

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

2001*	2002	2003	2004	2005	2006	2007
0	2	21	24	27	16	34

The word in context

IRAQ continues to flaunt its hostility toward America and to support terror. The Iraqi regime has plotted to develop anthrax, and nerve gas, and nuclear weapons for over a decade. This is a regime that has already used poison gas to murder thousands of its own citizens -- leaving the bodies of mothers huddled over their dead children. This is a regime that agreed to international inspections -- then kicked out the inspectors. This is a regime that has something to hide from the civilized world.

-- 2002 (Paragraph 20 of 67)

Compared with other words

	2001*	2002	2003	2004	2005	2006	2007
Iraq	-	2	21	24	27	16	34
Afghanistan	-	13	3	5	3	2	4
Economy(ic)	6	7	13	17	14	23	8
Insurance	2	-	1	6	1	3	14
Iraq/Iraqi(s)	-	2	21	24	27	16	34
Iran	-	2	3	1	3	6	5
Oil	-	1	-	-	-	3	9
Social Security	15	2	2	2	18	3	2

https://archive.nytimes.com/www.nytimes.com/ref/washington/20070123_STATEOFUNION.html?

86

THE WORDS THAT WERE USED

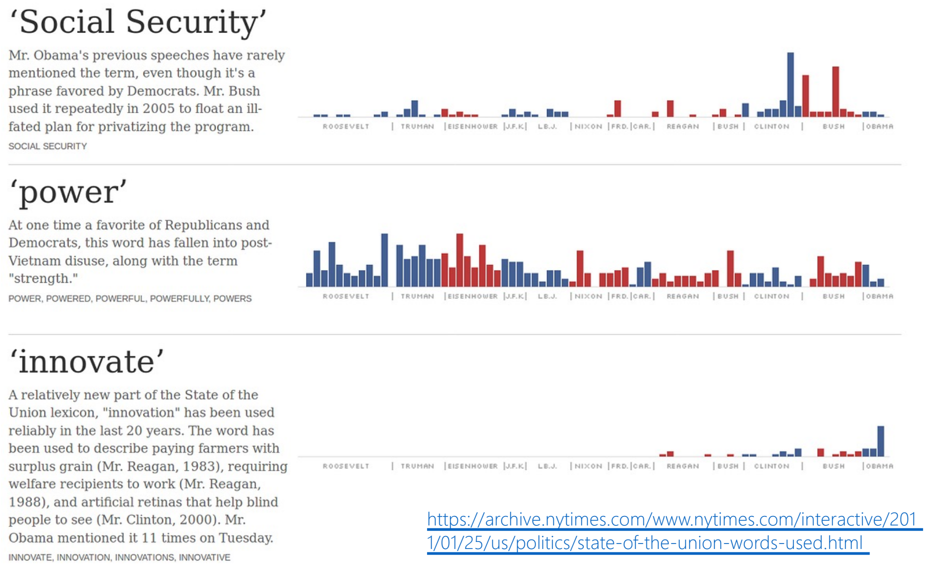
READ 2007 SPEECH | FEEDBACK

The 2007 State of the Union Address

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

https://archive.nytimes.com/www.nytimes.com/ref/washington/20070123_STATEOFUNION.html?

87



'Social Security'

Mr. Obama's previous speeches have rarely mentioned the term, even though it's a phrase favored by Democrats. Mr. Bush used it repeatedly in 2005 to float an ill-fated plan for privatizing the program.

SOCIAL SECURITY

'power'

At one time a favorite of Republicans and Democrats, this word has fallen into post-Vietnam disuse, along with the term "strength."

POWER, POWERED, POWERFUL, POWERFULLY, POWERS

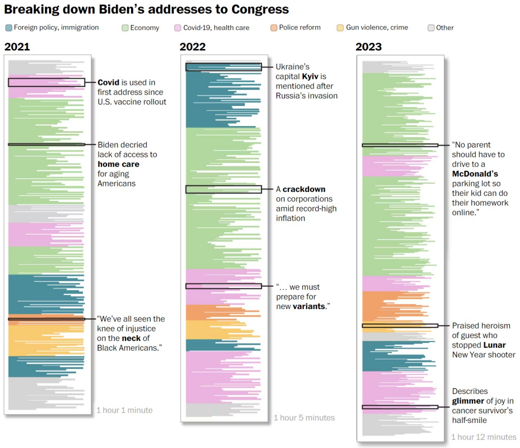
'innovate'

A relatively new part of the State of the Union lexicon, "innovation" has been used reliably in the last 20 years. The word has been used to describe paying farmers with surplus grain (Mr. Reagan, 1983), requiring welfare recipients to work (Mr. Reagan, 1988), and artificial retinas that help blind people to see (Mr. Clinton, 2000). Mr. Obama mentioned it 11 times on Tuesday.

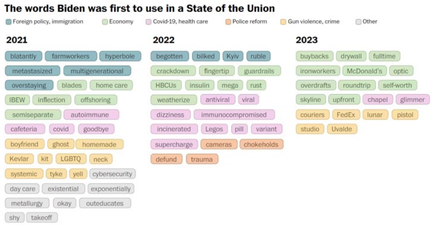
INNOVATE, INNOVATION, INNOVATIONS, INNOVATIVE

https://archive.nytimes.com/www.nytimes.com/interactive/2011/01/25/us/politics/state-of-the-union-words-used.html

88



Breaking down Biden's addresses to Congress



The words Biden was first to use in a State of the Union

https://www.washingtonpost.com/politics/interactive/2023/state-of-the-union-speech-history-quiz/

89

VISUALIZING CONVERSATIONS

90

MANY DIMESIONS TO CONSIDER

Dimensions

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

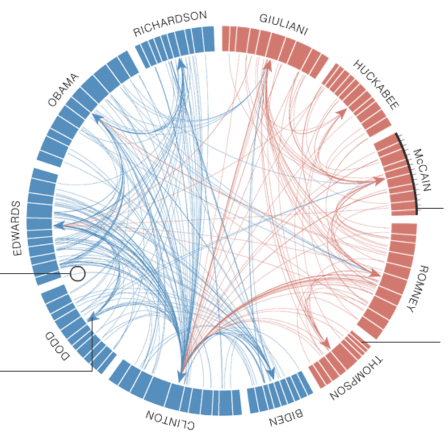
Cross-products

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

Naming Names

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

Roll over any candidate's name for details.



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

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

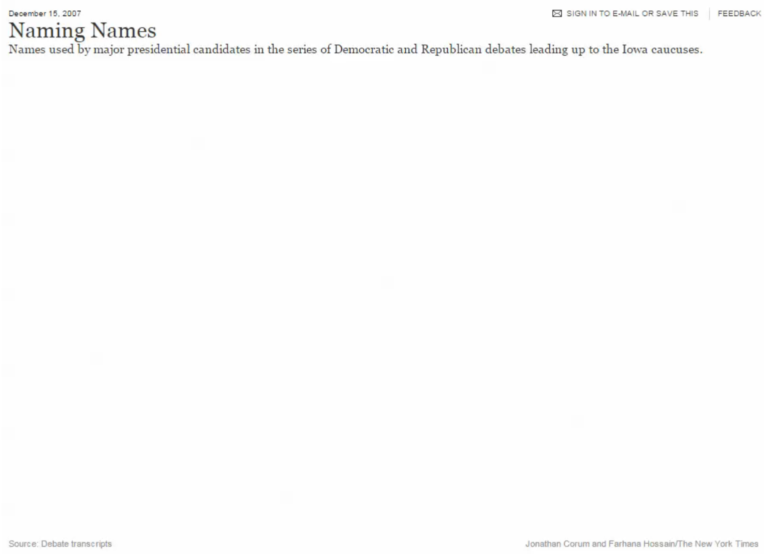
The length of each circle segment represents the total number of words spoken by the candidate during the debates. Each tick mark represents 1,000 words.

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

Source: Debate transcripts

Jonathan Corum and Farhana Hossain/The New York Times

http://www.nytimes.com/interactive/2007/12/15/us/politics/DEBATE.html



December 15, 2007

SIGN IN TO E-MAIL OR SAVE THIS | FEEDBACK

Naming Names

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

Source: Debate transcripts

Jonathan Corum and Farhana Hossain/The New York Times

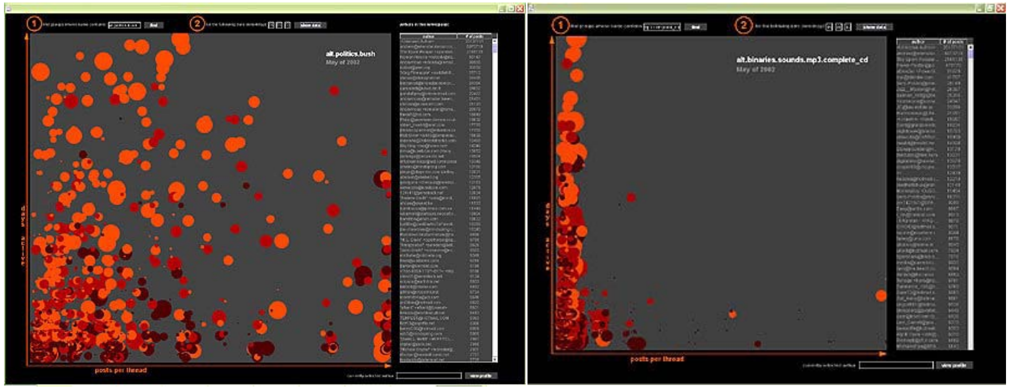
http://www.nytimes.com/interactive/2007/12/15/us/politics/DEBATE.html

93

USENET VISUALIZATION

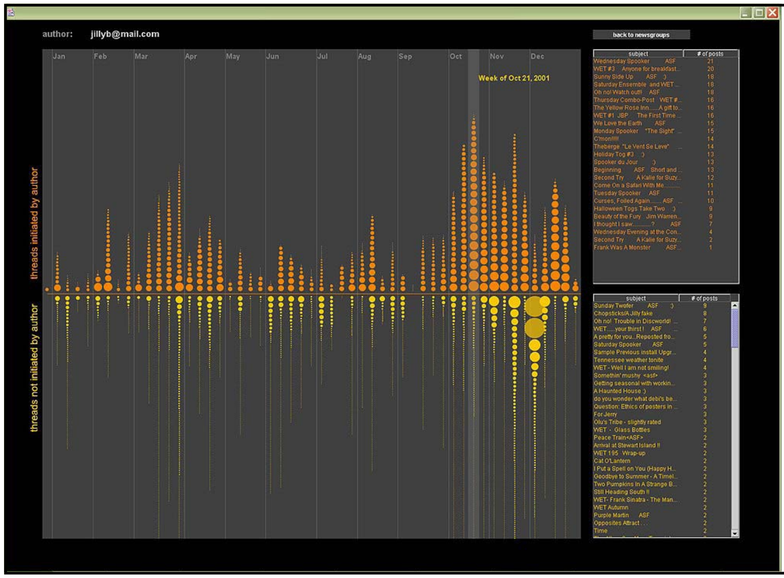
Show correspondence patterns in text forums

Initiate vs. reply; size and duration of discussion



[Viegas 2004]

94



author: jillyb@mail.com

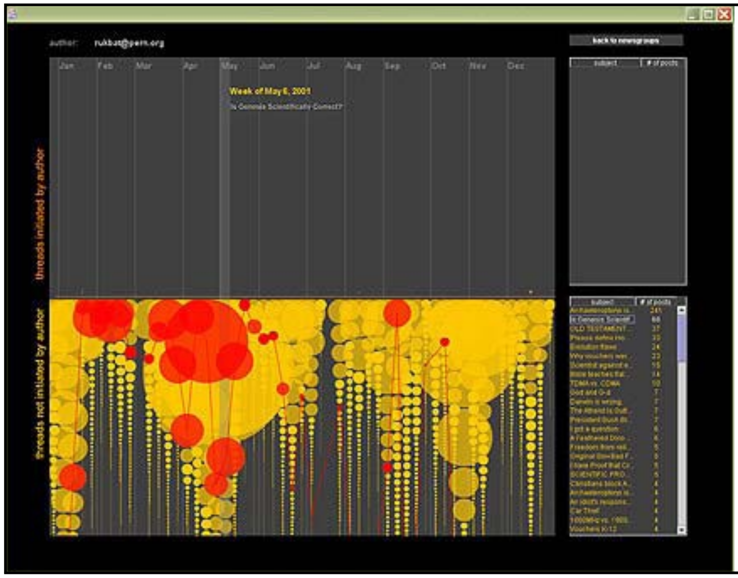
subject	# of posts
Wednesday Spooker ASF	21
WET #3 Anyone for breakfast...	20
Sunny Side Up ASF :)	18
Saturday Ensemble and WET ...	18
Oh no! Watch out!! ASF	18
Thursday Combo-Post WET #...	16
The Yellow Rose Inn......A gift to...	16
WET #1 JBP The First Time ...	16
We Love the Earth ASF	15
Monday Spooker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

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

Authorlines:
Shows activity of single author with respect to threads they initiated vs. those initiated by others

[Viegas 2004]

95



author: rukbat@pern.org

Is Genesis Scientifically Correct?

subject	# of posts
Archaeopteryx is...	241
Is Genesis Scientif...	66
OLD TESTAMENT...	37
Please define Ho...	33
Evolution flaws	24
Why vouchers won...	23
Scientist against e...	15
Bible teaches flat...	14
TDMA vs. CDMA	10
God and G-d	7
Darwin is wrong	7
The Atheist is Guilt...	7
President Bush di...	7
I got a question	6
A Feathered Dino...	6
Freedom from reli...	5
Original Sin/Bad F...	5
I have Proof that Cr...	5
SCIENTIFIC PRO...	5
Christians block K...	4
Archaeopteryx is...	4
An idiot's respons...	4
Car Thief	4
1000MHz vs. 1900...	4
Vouchers K-12	4

Authorlines:
Shows activity of single author with respect to threads they initiated vs. those initiated by others

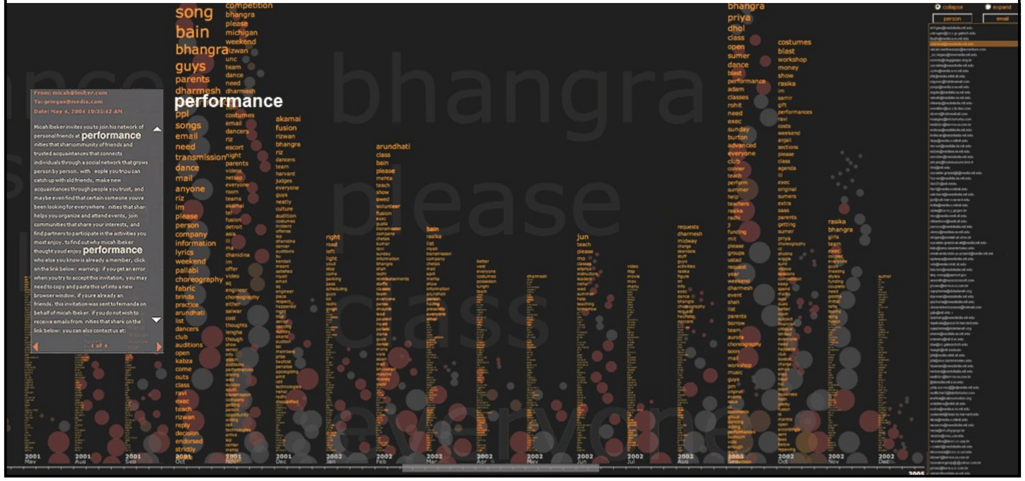
[Viegas 2004]

96

EMAIL VISUALIZATION

[Viegas 2006]

One person over time, TF.IDF weighted terms



SUMMARY

Text has many levels of possible visualizations

- Word, document, collection

Identifying descriptive words/keyphrases is critical

- e.g., TF, TF.IDF, regexp, …
- Domain dependent

Can go beyond standard charts and graphs but requires vigilance in design decisions (don't be seduced by possibilities)