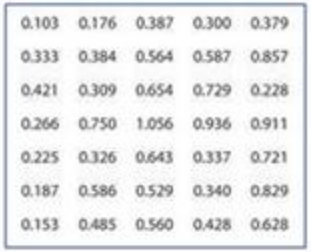
VISUALIZING TEXT

Petra Isenberg

RECAP

STRUCTURED DATA



0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

UNSTRUCTURED DATA

(TODAY)

VISUALIZING TEXT

TEXT IS DIFFERENT

COMMON

UNSTRUCTURED (MOSTLY)

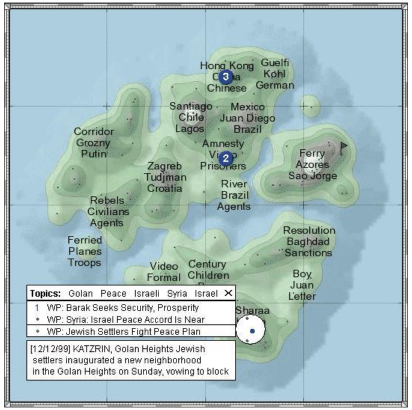
HIGH-DIMENSIONAL (10,000+)

BIG!

TEXT?

WHY

- To assist information retrieval
- To enable linguistic analysis
- To augment analytics on mixed data

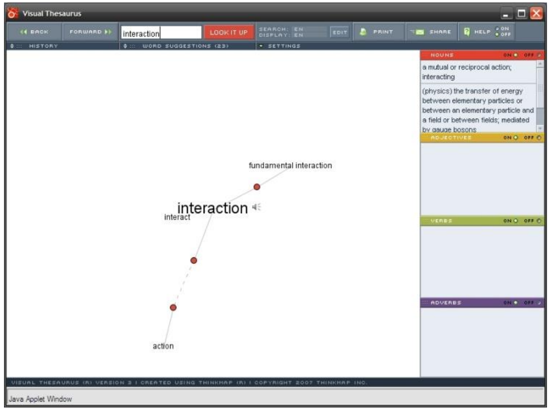


Topics: Golan Peace Israeli Syria Israel ✕

1 WP: Barak Seeks Security, Prosperity
- WP: Syria: Israel Peace Accord Is Near
- WP: Jewish Settlers Fight Peace Plan

[12/12/99] KATZRIN, Golan Heights Jewish
settlers inaugurated a new neighborhood
in the Golan Heights on Sunday, vowing to block

Themescape

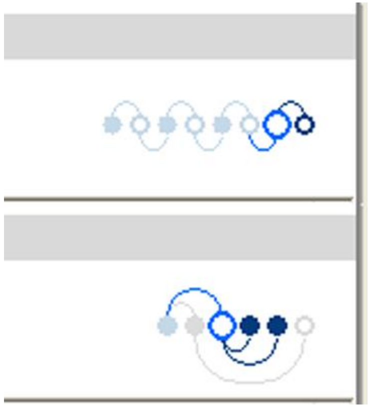


a mutual or reciprocal action; interacting

(physics) the transfer of energy between elementary particles or between an elementary particle and a field or between fields; mediated by gauge bosons

VISUAL THESAURUS (R) VERSION 3 | CREATED USING THINKMAP (R) | COPYRIGHT 2007 THINKMAP INC.

Visual Thesaurus



Thread Arcs

WHY

UNDERSTANDING: GET THE “GIST” OF A DOCUMENT

GROUPING: CLUSTER FOR OVERVIEW OR CLASSIFICATION

COMPARE: COMPARE DOCUMENT COLLECTIONS, OR INSPECT EVOLUTION OF COLLECTION OVER TIME

CORRELATE: COMPARE PATTERNS IN TEXT TO THOSE IN OTHER DATA, E.G., CORRELATE WITH SOCIAL NETWORK

WHAT IS TEXT

DOCUMENTS

- ARTICLES, BOOKS AND NOVELS
- COMPUTER PROGRAMS
- E-MAILS, WEB PAGES, BLOGS
- TAGS, COMMENTS

COLLECTION OF DOCUMENTS

- MESSAGES (E-MAIL, BLOGS, TAGS, COMMENTS)
- SOCIAL NETWORKS (PERSONAL PROFILES)
- ACADEMIC COLLABORATIONS (PUBLICATIONS)
- EVEN WHOLE LIBRARIES, WEBSITES, SOCIAL NETWORKS

DIFFICULT DATA

- Too much data – what to use?
  - Millions of blog posts,
  - Hundreds of thousands of news stories,
  - 183 billion emails,
  - … per day
- Data is noisy:
  - 70-72% of email is spam
  - Text contains section headings, figure captions, and direct quotes
  - ….

ONCE YOU HAVE THE DATA…

- Most meaning comes from our minds and common understanding.

- “How much is that doggy in the window?”
  - how much: social system of barter and trade (not the size of the dog)
  - “doggy” implies childlike, plaintive, probably cannot do the purchasing on their own
  - “in the window” implies behind a store window, not really inside a window, requires notion of window shopping

(Hearst, 2006)

LANGUAGE IS AMBIGUOUS

- Words and phrases can have many meanings, determined by context and world knowledge

- Interesting language is often figurative:
  - *America is a melting pot* (metaphor)
  - *Busy as a bee* (simile)
  - *Opportunity knocked on the door* (personification)
  - *You could have knocked me over with a feather* (hyperbole)

LANGUAGE IS AMBIGUOUS

"I can't tell you how much I enjoyed meeting your husband."

(William Empson, *Seven Types of Ambiguity*, 1947)

"Brave men run in my family."

(Bob Hope as "Painless" Peter Potter in *The Paleface*, 1948)

VISUAL CONSIDERATIONS

Supporters of Martin, who has been jailed without trial for more than two years, are calling on Prime Minister Stephen Harper to ask Mexican president Felipe Calderon to release Martin text is not preattentive under a section of the Mexican constitution that allows the government to expel undesirables from the country. Martin's supporters believe she has no chance of a fair trial in Mexico. Neither does Waage.

VISUAL CONSIDERATIONS

Supporters of Martin, who has been jailed without trial for more than two years, are calling on Prime Minister Stephen Harper to ask Mexican president Felipe Calderon to release Martin text is not preattentive under a section of the Mexican constitution that allows the government to expel undesirables from the country. Martin's supporters believe she has no chance of a fair trial in Mexico. Neither does Waage.

VISUAL CONSIDERATIONS

Text readability is dependent on size, orientation, font, clutter…

VISUAL CONSIDERATIONS

- Text readability is dependent on size, orientation, font, clutter...
- More likely to need large amounts of text in language visualization

VISUALIZING LANGUAGE IS ALSO EASY!

- SO much data available for analysis
- (Mostly) readily computer readable
- Simple techniques can give instant summaries

OUTLINE

- TEXT AS DATA
- VISUALIZING DOCUMENT CONTENT
- EVOLVING DOCUMENTS
- DOCUMENT COLLECTIONS

TEXT AS DATA

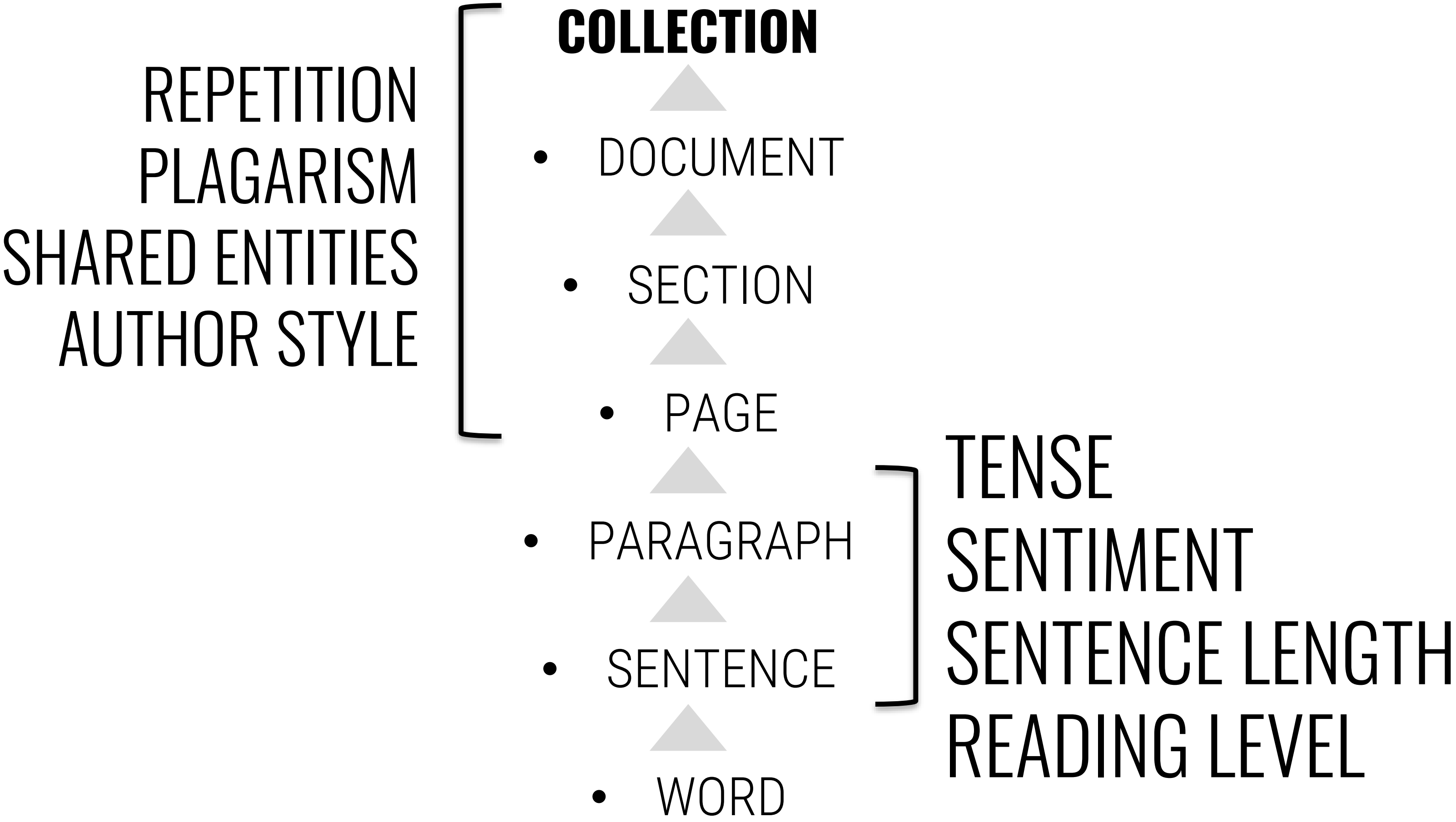
Words are the basic unit of data.

WORD-LEVEL ATTRIBUTES

- WORD LENGTH
- PART OF SPEECH (NOUN, VERB, ADJECTIVE, ETC.)
- FORMAT (*ITALIC*, UNDERLINE, ETC.)
- LANGUAGE (ENGLISH? LATIN? JAPANESE?)
- FREQUENCY / DIFFICULTY (IS IT COMMON?)
- SENTIMENT (POSITIVE OR NEGATIVE CONNOTATION)
- SYNONYMS / ANTONYMS / ETYMOLOGY (OTHER MEANINGS? ROOTS?)
- ENTITIES (e.g. “Calgary”, “Obama”, “Telus”)

- … AND MANY MORE

AGGREGATION



COLLECTION

- DOCUMENT
- SECTION
- PAGE
- PARAGRAPH
- SENTENCE
- WORD

REPETITION
PLAGARISM
SHARED ENTITIES
AUTHOR STYLE

TENSE
SENTIMENT
SENTENCE LENGTH
READING LEVEL

LINGUISTIC METHODS

- Word Counting
- Word Scoring
- Stemming
- Stop Word Removal
- Part of Speech Tagging
- Parsing
- Word Sense Disambiguation
- Named Entity Recognition
- Semantic Categorization
- Sentiment Analysis
- Topic Modeling (some caveats)

WHAT ABOUT THESE WORDS?

automate
automates
automatic
automation

→ automat

~~a, an, the, to, …~~

“New York”
“Ban Ki-moon”
“Manchester United”

STEMMING

- Reduce words to their ‘stems’ by removing endings (morphology)
  - running -> run
  - runs -> run

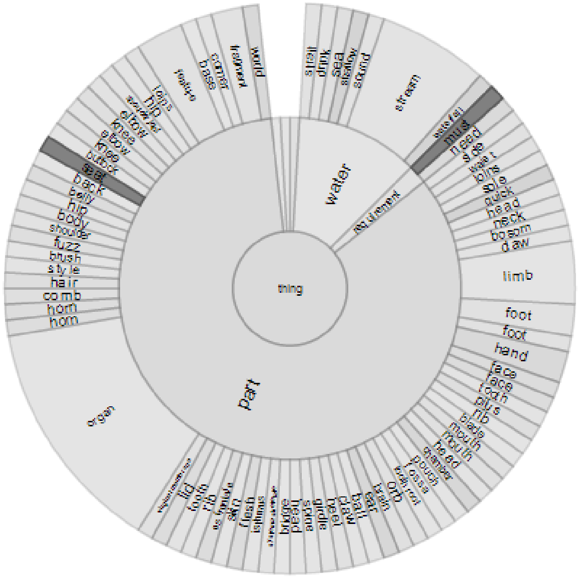
- A good way to increase signal and reduce fracturing of the corpus if there aren’t many words.
- Note: Keep the original words somewhere! Also keep the case if you choose to lowercase the word; you never know when you’ll need this data

STOP WORD REMOVAL

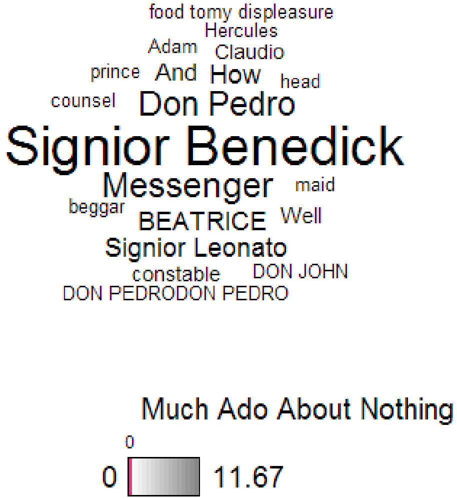
- Common words such as “and”, “the”, “I” are removed from view to highlight content words
- Domain specific stop words, e.g. in legal domain:
  - Court, attorney, honour, plaintiff, etc.
- Caution! These words have been shown to be useful for stylistic analysis! When working with text corpora, KEEP EVERYTHING.

NAMED ENTITY RECOGNITION

- What are the people, places in the text?
- Use NLTK – it's very good at this.



http://vialab.science.uoit.ca/docuburst



TEXT PROCESSING

TOKENIZATION: SEGMENT TEXT INTO TERMS

ENTITIES? “SAN FRANCISCO”, “O’CONNOR”, “U.S.A.”

REMOVE STOP WORDS? “A”, “AN”, “THE”, “TO”, “BE”

N-GRAMS? CAN TAKE WORDS IN 2-WORD GROUPS (BI-GRAMS), 3-WORD (TRI-GRAMS), ETC.

STEMMING: GROUP TOGETHER DIFFERENT FORMS

ROOTS: VISUALIZATION(S), VISUALIZE(S), VISUALLY ➞ VISUAL

LEMMATIZATION: GOES, WENT, GONE ➞ GO

FOR VISUALIZATION, SOMETIMES NEED TO REVERSE STEMMING FOR LABELS

SIMPLE SOLUTION: MAP FROM STEM TO THE MOST FREQUENT WORD

RESULT: ORDERED STREAM OF TERMS

TEXT PROCESSING

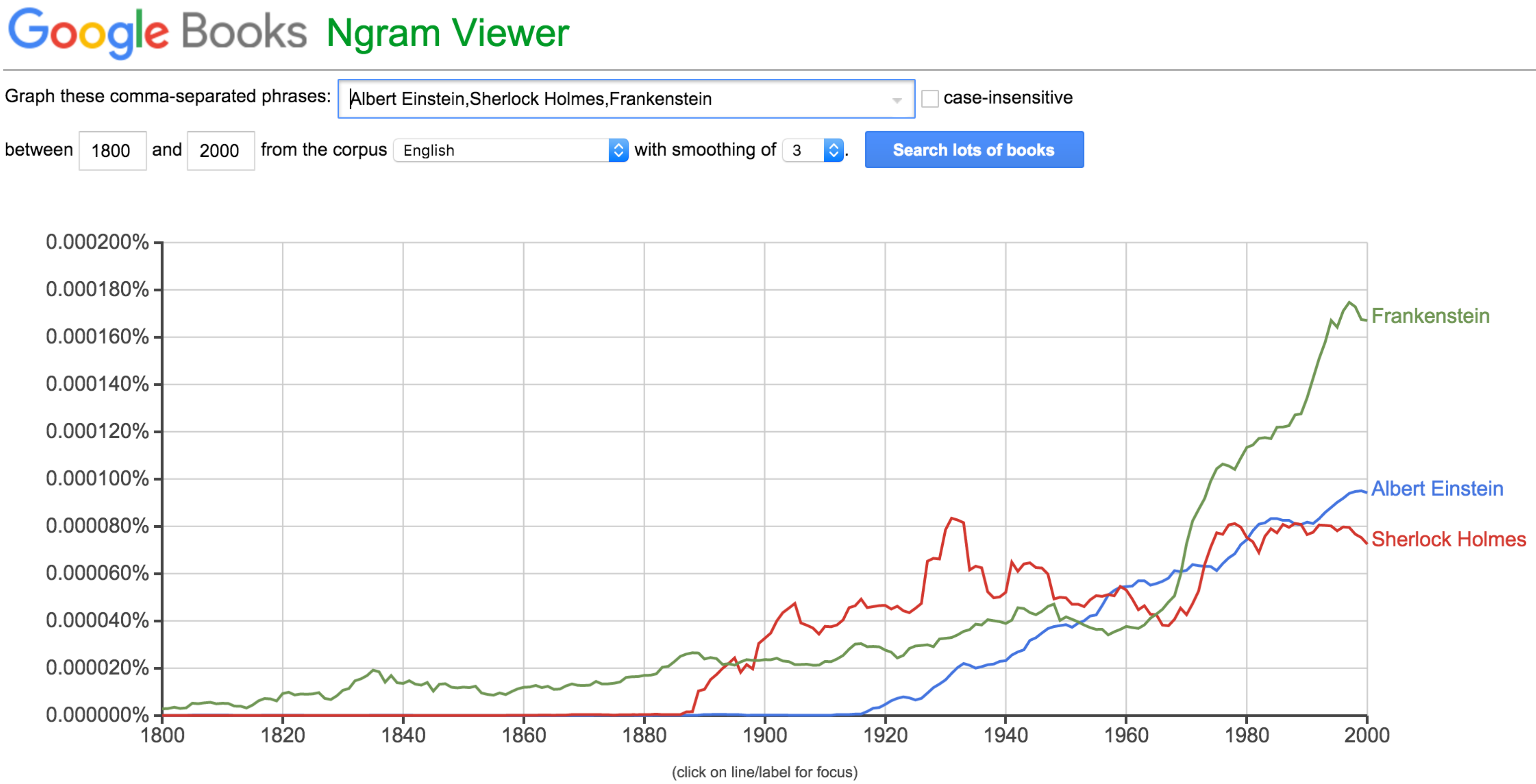
“The quick brown fox jumps over the lazy dog.”

TOKENIZE (N=1)
[The], [quick], [brown], [fox], [jumps], [over], [the], [lazy], [dog].

TOKENIZE (N=1), REMOVE STOPWORDS, STEM
[quick], [brown], [fox], [jump], [over], [lazy], [dog]

TOKENIZE (N=2)
[the quick], [quick brown], [brown fox], [fox jumps], [jumps over], [over the]…

TOKENIZE (N=5)
[the quick brown fox jumps], [quick brown fox jumps over], [brown fox jumps over …

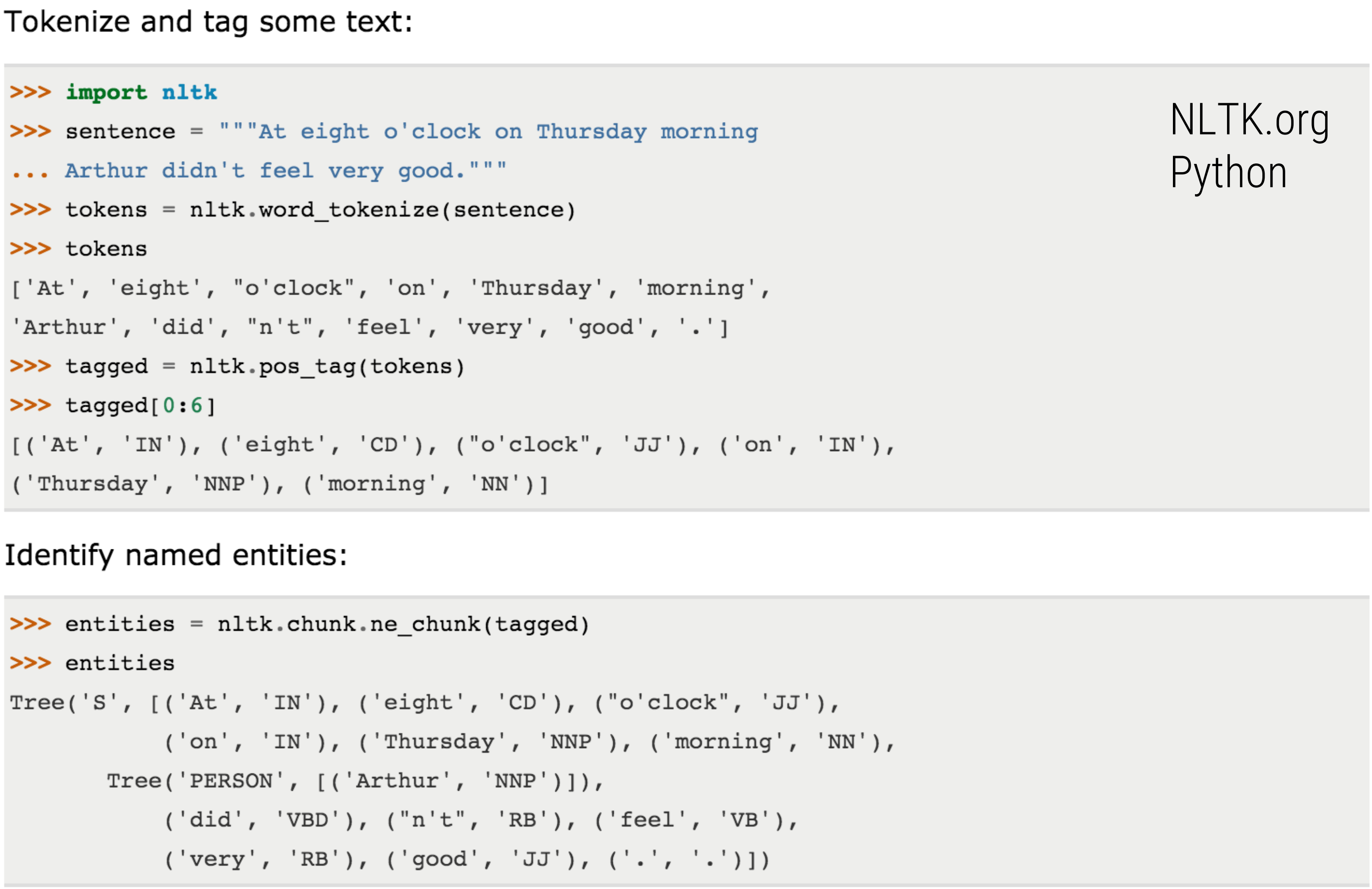


Google Books Ngram Viewer

Graph these comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein □ case-insensitive

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

NLTK (NATURAL LANGUAGE TOOLKIT)



Tokenize and tag some text:

```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',
'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
>>> tagged = nltk.pos_tag(tokens)
>>> tagged[0:6]
[('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'), ('on', 'IN'),
('Thursday', 'NNP'), ('morning', 'NN')]
```

Identify named entities:

```
>>> entities = nltk.chunk.ne_chunk(tagged)
>>> entities
Tree('S', [('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'),
           ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN'),
       Tree('PERSON', [('Arthur', 'NNP')]),
           ('did', 'VBD'), ("n't", 'RB'), ('feel', 'VB'),
           ('very', 'RB'), ('good', 'JJ'), ('.', '.')])
```

NLTK.org
Python

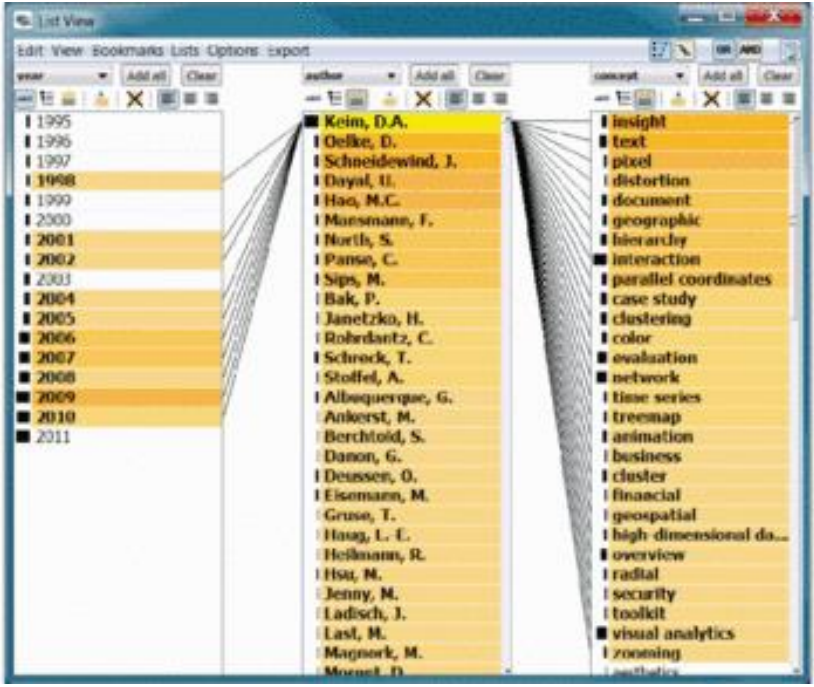
NAMED ENTITY RECOGNITION

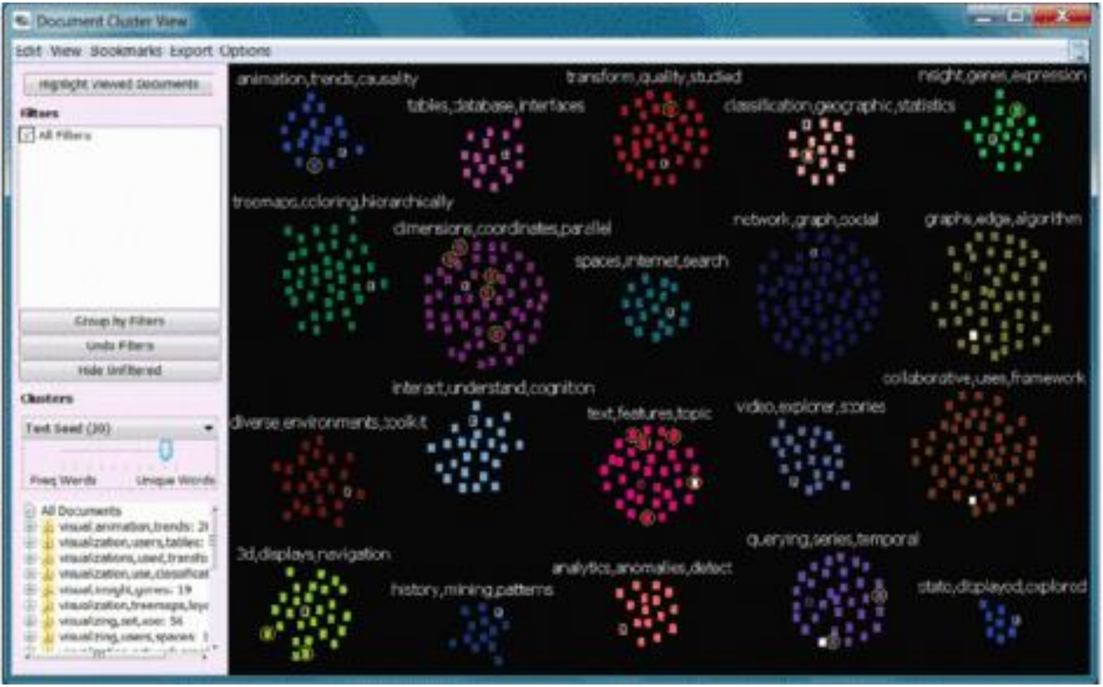
IDENTIFY AND CLASSIFY NAMED ENTITIES IN TEXT:

JOHN SMITH IS A PERSON
SOVIET UNION IS A COUNTRY
2500 UNIVERSITY DR IS AN ADDRESS
(555) 867-5309 IS A PHONE NUMBER

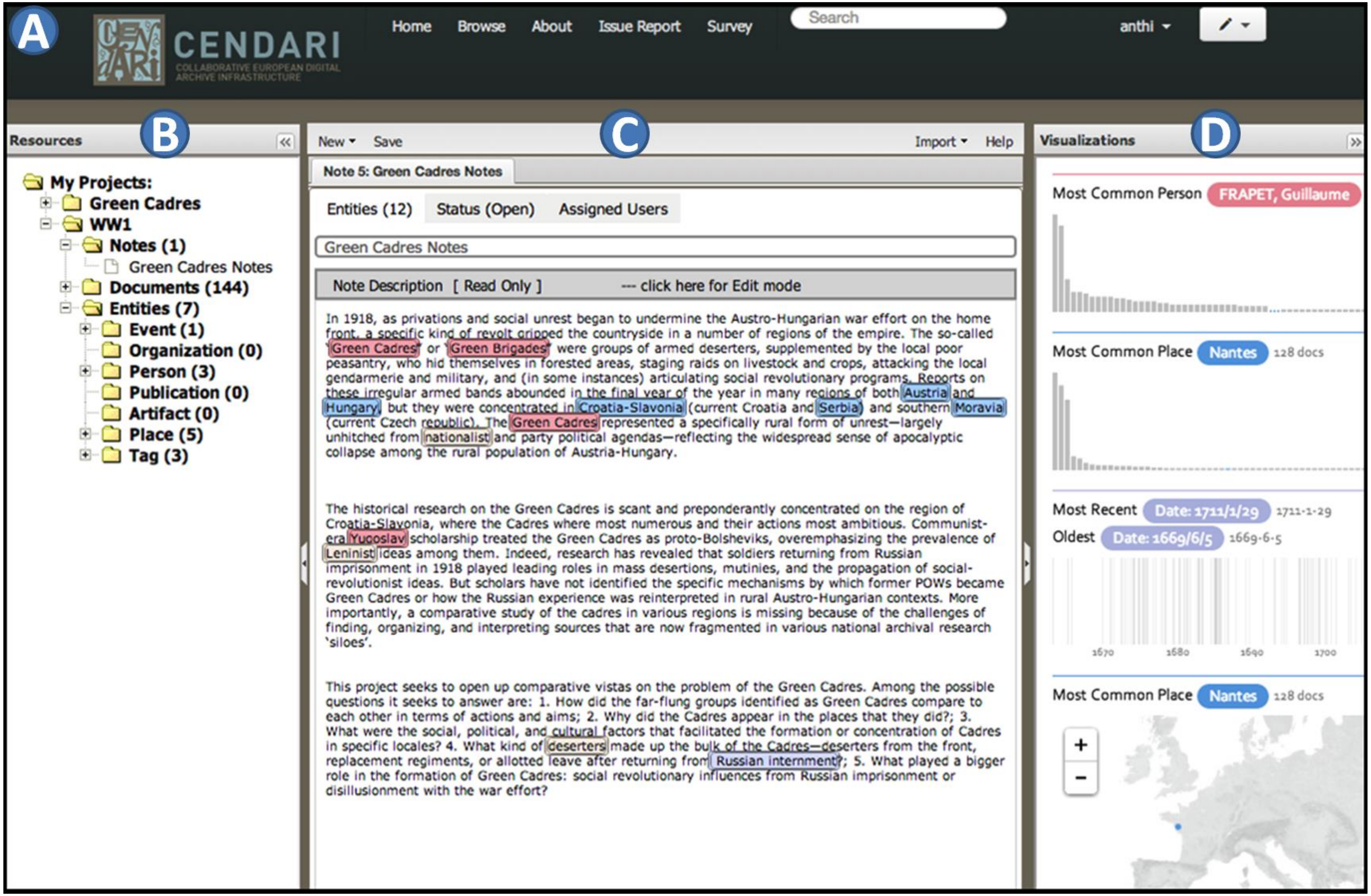
ENTITY RELATIONS: HOW DO THE ENTITIES RELATE?

DO THEY CO-OCCUR IN A DOCUMENT? IN A SENTENCE?


List View
Edit View Bookmarks Lists Options Export
year
Add all
Clear
author
Add all
Clear
concept
Add all
Clear
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Keim, D.A.
Oelke, D.
Schneidewind, J.
Dayal, U.
Hao, M.C.
Mansmann, F.
North, S.
Panse, C.
Sips, M.
Bak, P.
Janetzko, H.
Rohrdantz, C.
Schreck, T.
Stoffel, A.
Albuquerque, G.
Ankerst, M.
Berchtold, S.
Danon, G.
Deussen, O.
Eisemann, M.
Gruse, T.
Haag, L. L.
Heilmann, R.
Hsu, M.
Jenny, M.
Ladisch, J.
Last, M.
Magnork, M.
insight
text
pixel
distortion
document
geographic
hierarchy
interaction
parallel coordinates
case study
clustering
color
evaluation
network
time series
treemap
animation
business
cluster
financial
geospatial
high dimensional da...
overview
radial
security
toolkit
visual analytics
zooming


Document Cluster View
Edit View Bookmarks Export Options
Highlight Viewed Documents
Filters
All Filters
Group by Filters
Undo Filters
Hide Unfiltered
Clusters
Text Seed (30)
Freq Words
Unique Words
All Documents
visual,animation,trends: 21
visualization,users,tables:
visualizations,used,transfo
visualization,use,classificat
visual,insight,genes: 19
visualization,treemaps,layo
visualizing,set,user: 56
visualizing,users,spaces: 1
animation,trends,causality
tables,database,interfaces
transform,quality,studied
classification,geographic,statistics
insight,genes,expression
treemaps,coloring,hierarchically
dimensions,coordinates,parallel
spaces,internet,search
network,graph,social
graphs,edge,algorithm
interact,understand,cognition
diverse,environments,toolkit
text,features,topic
video,explorer,stories
collaborative,uses,framework
3d,displays,navigation
history,mining,patterns
analytics,anomalies,detect
querying,series,temporal
static,displayed,explored

JIGSAW



In 1918, as privations and social unrest began to undermine the Austro-Hungarian war effort on the home front, a specific kind of revolt gripped the countryside in a number of regions of the empire. The so-called 'Green Cadres' or 'Green Brigades' were groups of armed deserters, supplemented by the local poor peasantry, who hid themselves in forested areas, staging raids on livestock and crops, attacking the local gendarmerie and military, and (in some instances) articulating social revolutionary programs. Reports on these irregular armed bands abounded in the final year of the year in many regions of both Austria and Hungary, but they were concentrated in Croatia-Slavonia (current Croatia and Serbia) and southern Moravia (current Czech republic). The Green Cadres represented a specifically rural form of unrest—largely unhitched from nationalist and party political agendas—reflecting the widespread sense of apocalyptic collapse among the rural population of Austria-Hungary.

The historical research on the Green Cadres is scant and preponderantly concentrated on the region of Croatia-Slavonia, where the Cadres where most numerous and their actions most ambitious. Communist-era Yugoslav scholarship treated the Green Cadres as proto-Bolsheviks, overemphasizing the prevalence of Leninist ideas among them. Indeed, research has revealed that soldiers returning from Russian imprisonment in 1918 played leading roles in mass desertions, mutinies, and the propagation of social-revolutionist ideas. But scholars have not identified the specific mechanisms by which former POWs became Green Cadres or how the Russian experience was reinterpreted in rural Austro-Hungarian contexts. More importantly, a comparative study of the cadres in various regions is missing because of the challenges of finding, organizing, and interpreting sources that are now fragmented in various national archival research 'siloes'.

This project seeks to open up comparative vistas on the problem of the Green Cadres. Among the possible questions it seeks to answer are: 1. How did the far-flung groups identified as Green Cadres compare to each other in terms of actions and aims; 2. Why did the Cadres appear in the places that they did?; 3. What were the social, political, and cultural factors that facilitated the formation or concentration of Cadres in specific locales? 4. What kind of deserters made up the bulk of the Cadres—deserters from the front, replacement regiments, or allotted leave after returning from Russian internment?; 5. What played a bigger role in the formation of Green Cadres: social revolutionary influences from Russian imprisonment or disillusionment with the war effort?

CENDARI NOTE-TAKING ENVIRONMENT 2015

DOCUMENT CONTENT

BUT FIRST **SOME SKETCHING**

SKETCHING: **VISUALIZE**

IMAGINE YOU HAVE A MASTER'S THESIS IN FROM OF YOU:

YEAR

AUTHOR

TITLE

KEYWORDS

REFERENCES

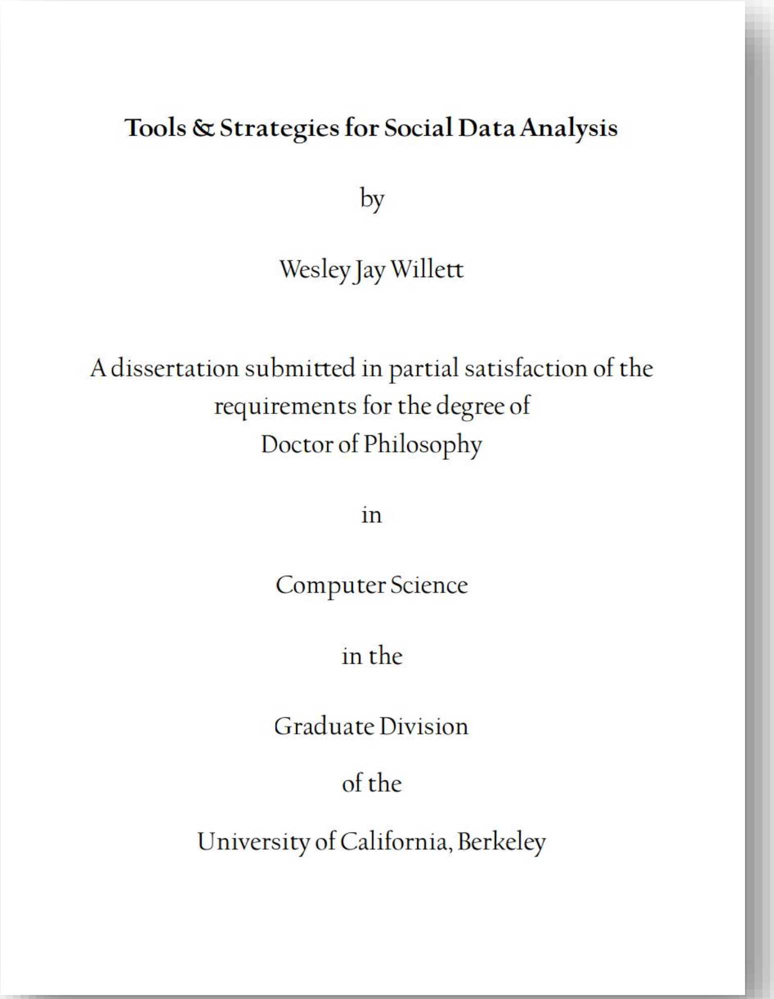
ABSTRACT TEXT

TASK:

1) VISUALIZE THE MOST IMPORTANT CONTENT FROM A SINGLE THESIS.

(~10 MINUTES)

EXAMPLE



Tools & Strategies for Social Data Analysis

by

Wesley Jay Willett

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

Graduate Division

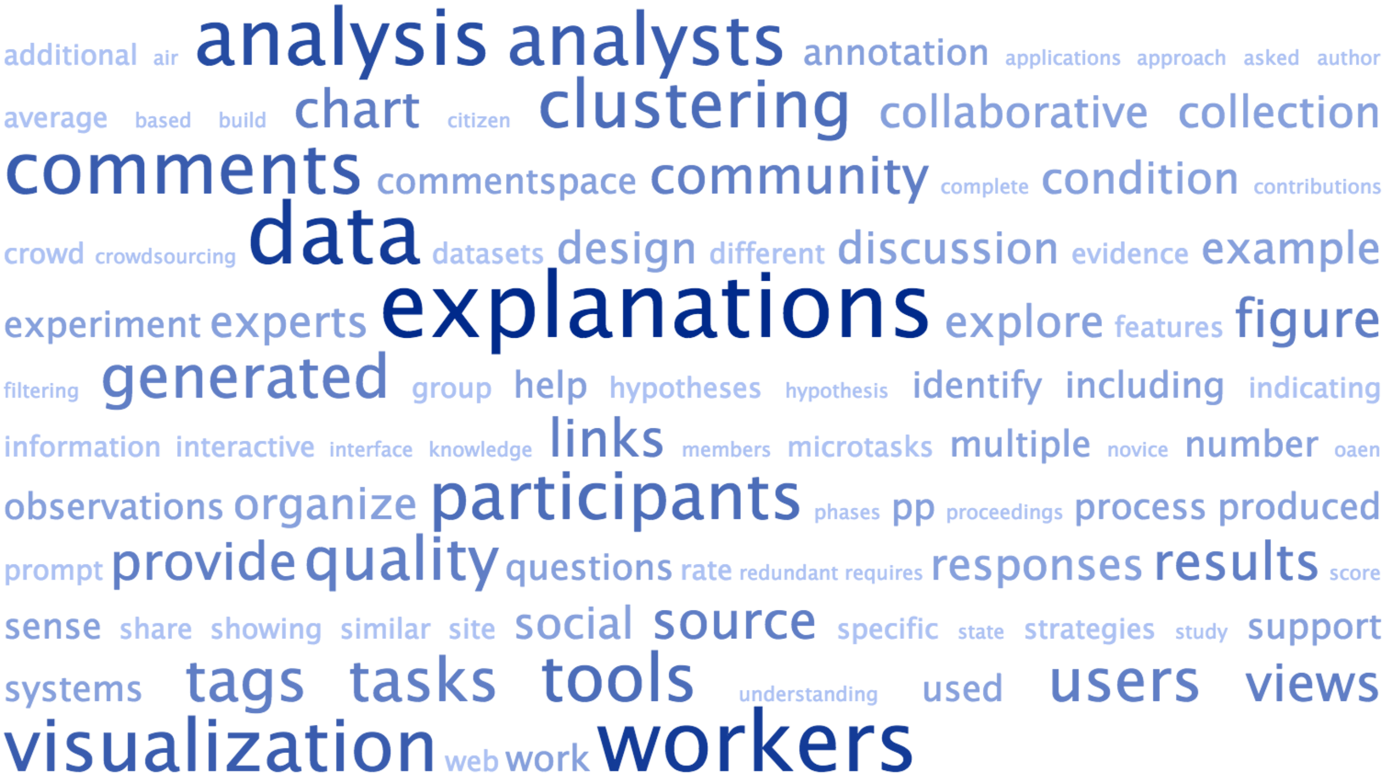
of the

University of California, Berkeley

THESIS WESLEY WILLETT

TAG CLOUDS

WORD COUNT

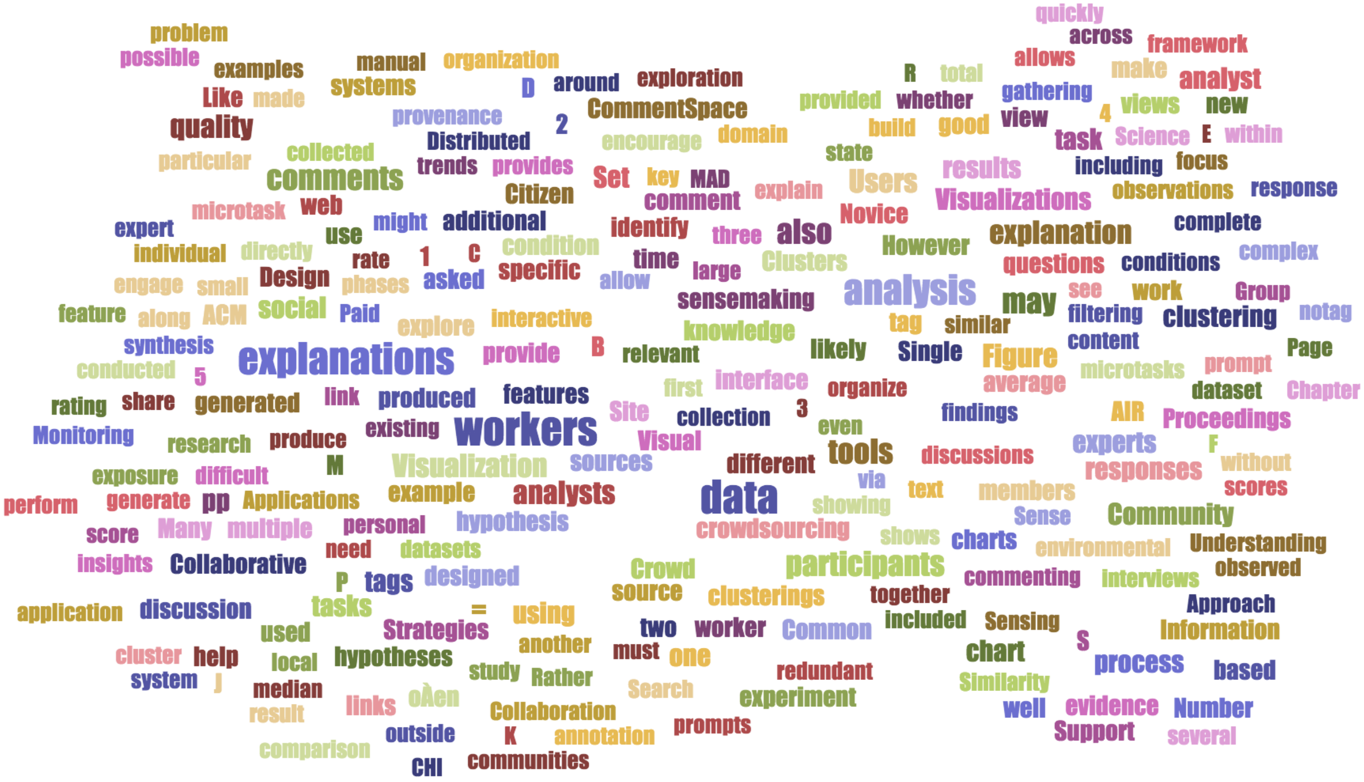


http://tagcrowd.com/

THESIS WESLEY WILLETT

TAG CLOUDS

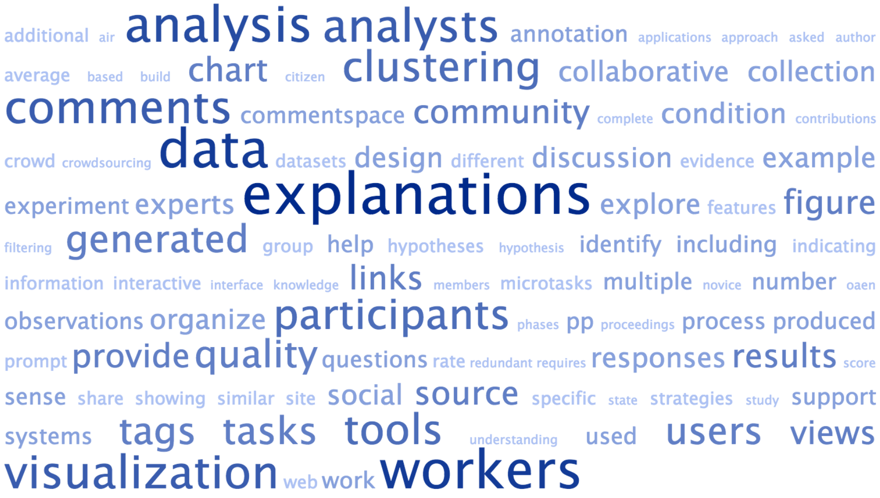
WORD COUNT

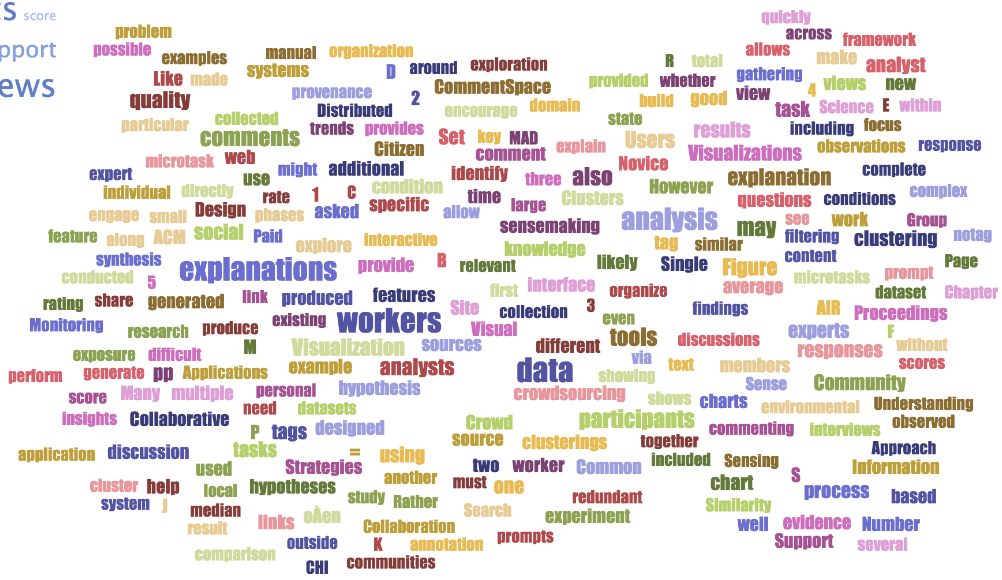


www.jasondavies.com/wordcloud/

THESIS WESLEY WILLETT

WHAT'S PROBLEMS DO YOU SEE WITH TAG CLOUDS?





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)

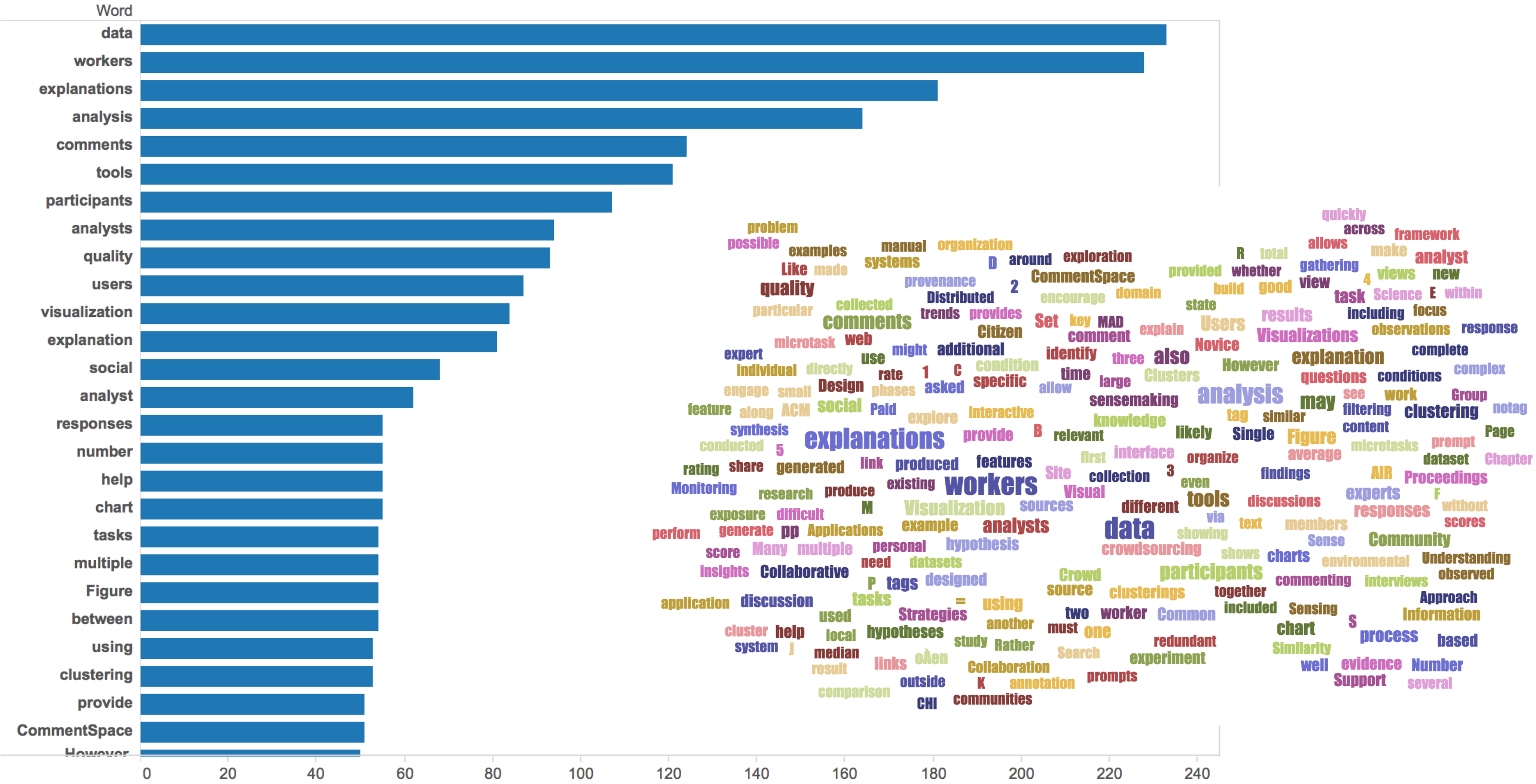
MAY NOT FACILITATE COMPARISON (UNSTABLE LAYOUT)

ORDER USUALLY MEANINGLESS (USUALLY ALPHABETICAL OR RANDOM)

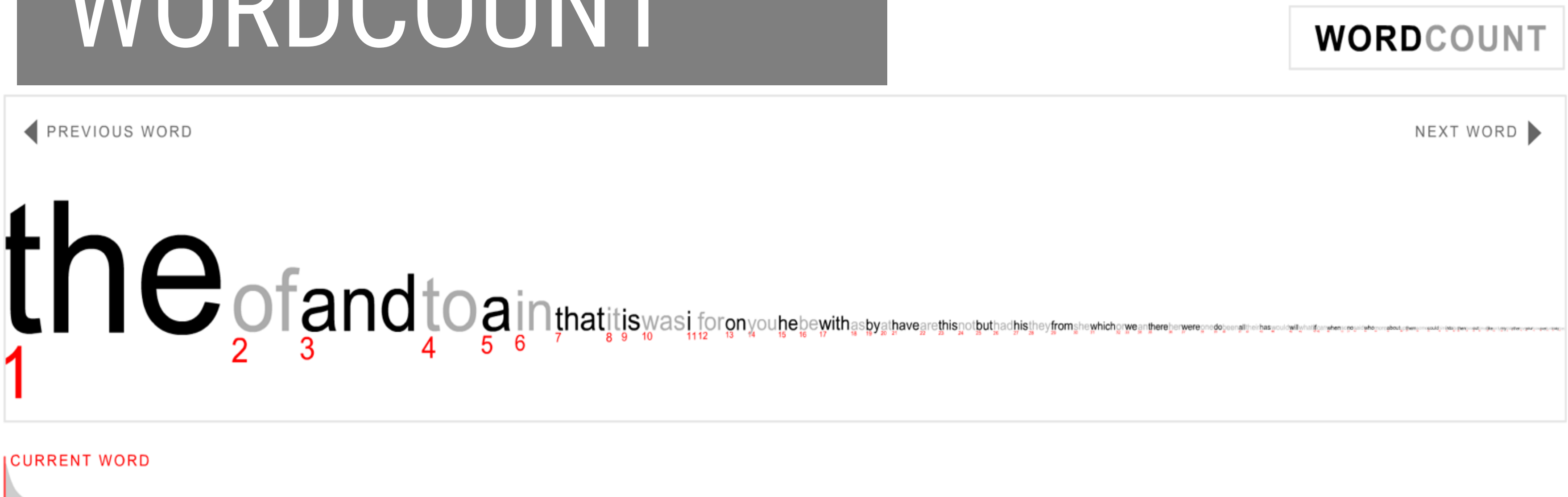
TERM FREQUENCY MAY NOT BE MEANINGFUL

DOES NOT SHOW THE STRUCTURE OF THE TEXT

WORD COUNTS


Word
data
workers
explanations
analysis
comments
tools
participants
analysts
quality
users
visualization
explanation
social
analyst
responses
number
help
chart
tasks
multiple
Figure
between
using
clustering
provide
CommentSpace
However
0
20
40
60
80
100
120
140
160
180
200
220
240
quickly
across framework
problem
possible examples manual organization allows make analyst
systems D around exploration R total gathering
Like made provenance 2 CommentSpace provided whether 4 views new
quality collected Distributed encourage domain build good view task Science E within
particular comments trends provides state results including focus
Set key MAD Users
Citizen comment explain Visualizations observations response
microtask web Novice
expert use might additional identify three also explanation complete
individual directly condition However
rate 1 C time Clusters questions conditions complex
engage small Design phases asked specific allow large
feature along ACM social Paid explore interactive sensemaking analysis may see work Group
filtering clustering notag
synthesis knowledge tag similar
conducted 5 explanations provide B relevant likely Single Figure content Page
first interface organize average microtasks prompt
rating share generated link produced features 3 dataset Chapter
Site collection findings AIR Proceedings
Monitoring research produce existing workers Visual even experts F
exposure difficult M Visualization sources different tools discussions responses without
perform generate pp Applications example analysts data via text members scores
showing
score Many multiple personal hypothesis Sense Community
crowdsourcing shows charts environmental Understanding
insights Collaborative need datasets participants commenting interviews observed
P tags designed Crowd
source clusterings together Approach
application discussion = using
tasks
used Strategies two worker Common included Sensing Information
cluster help another must one S
system local hypotheses study Rather chart process based
median redundant Similarity
J oÀen Search experiment well evidence Number
result links Collaboration prompts Support several
outside K annotation
comparison CHI communities

WORDCOUNT


WORDCOUNT
PREVIOUS WORD
NEXT WORD
the of and to a in that it is was i for on you he be with as by at have are this not but had his they from she which or we an there her were
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
CURRENT WORD

FIND WORD:

BY RANK:

REQUESTED WORD: THE

RANK: 1

ARCHIVE

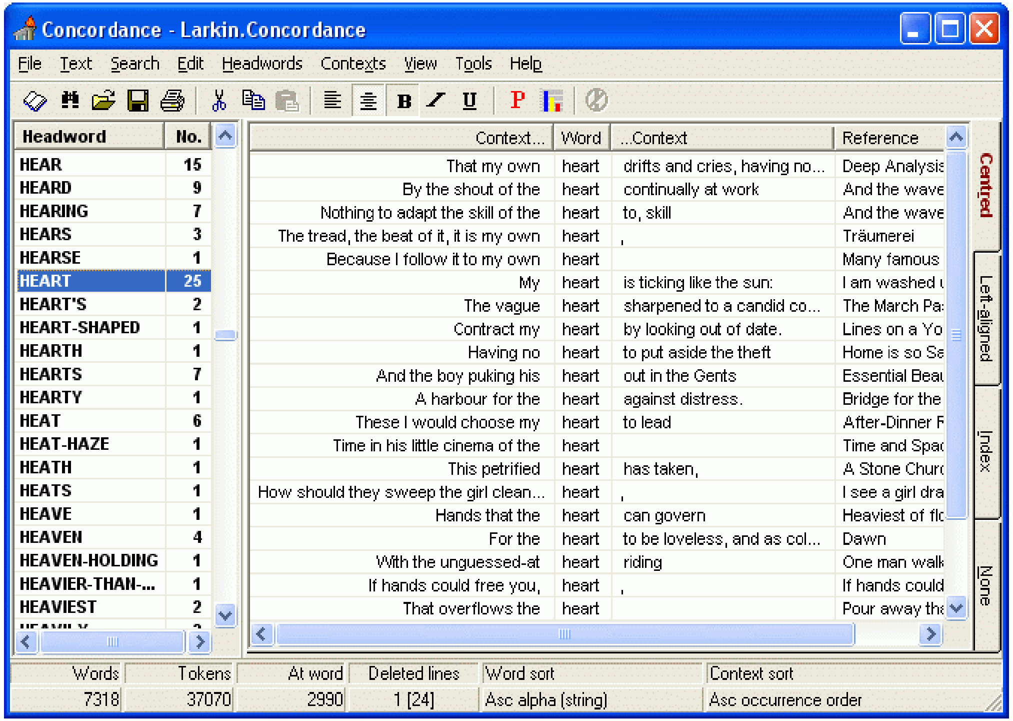
COUNT

http://wordcount.org

JONATHAN HARRIS

CONCORDANCE

WHAT IS THE COMMON LOCAL CONTEXT OF A TERM?



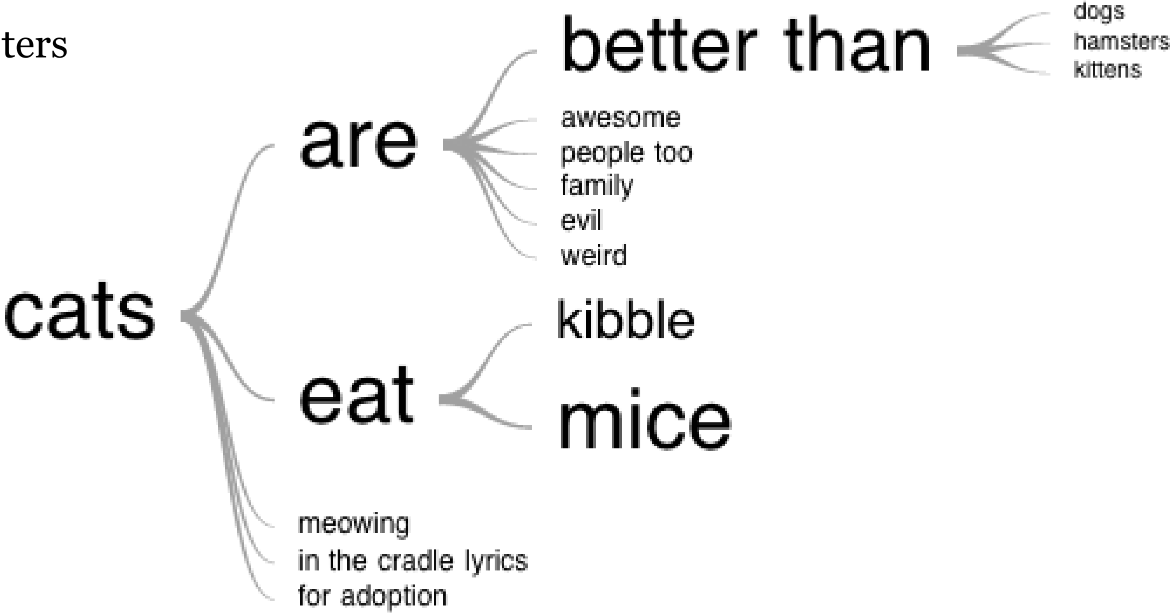
Headword	No.
HEAR	15
HEARD	9
HEARING	7
HEARS	3
HEARSE	1
HEART	25
HEART'S	2
HEART-SHAPED	1
HEARTH	1
HEARTS	7
HEARTY	1
HEAT	6
HEAT-HAZE	1
HEATH	1
HEATS	1
HEAVE	1
HEAVEN	4
HEAVEN-HOLDING	1
HEAVIER-THAN-...	1
HEAVIEST	2

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 u
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 the

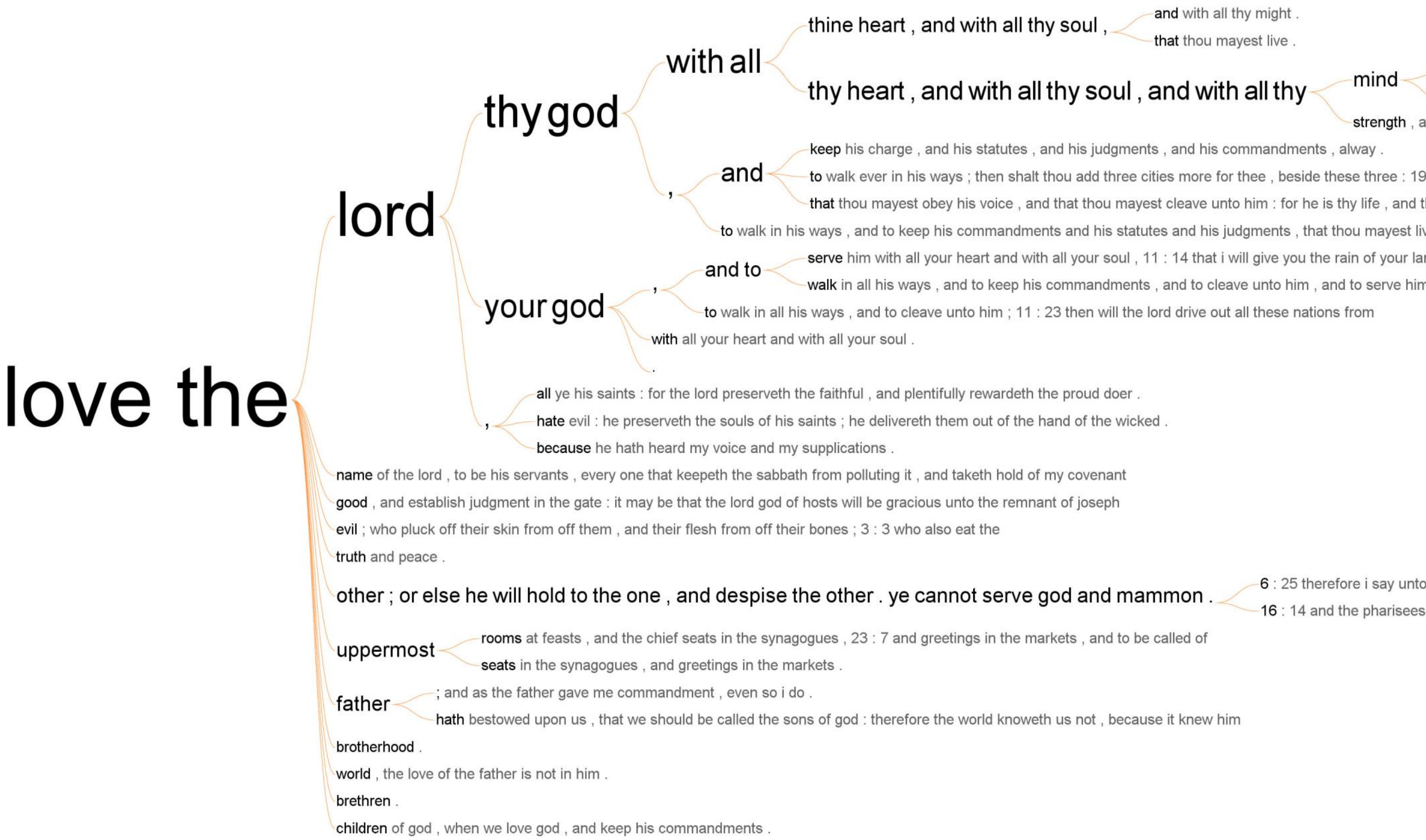
Words	Tokens	At word	Deleted lines	Word sort	Context sort
7318	37070	2990	1 [24]	Asc alpha (string)	Asc occurrence order

WORD TREES

- cats are better than dogs
- cats eat kibble
- cats are better than hamsters
- cats are awesome
- cats are people too
- cats eat mice
- cats meowing
- cats in the cradle
- cats eat mice
- cats in the cradle lyrics
- cats eat kibble
- cats for adoption
- cats are family
- cats eat mice
- cats are better than kittens
- cats are evil
- cats are weird
- cats eat mice



WATTENBERG & VIÉGAS 2008



31
hits

love the

lord

thy god

with all

thine heart , and with all thy soul ,

and with all thy might .

that thou mayest live .

thy heart , and with all thy soul , and with all thy

mind

strength , a

, and

keep his charge , and his statutes , and his judgments , and his commandments , alway .

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 t

to walk in his ways , and to keep his commandments and his statutes and his judgments , that thou mayest liv

your god

, and to

serve him with all your heart and with all your soul , 11 : 14 that i will give you the rain of your la

walk in all his ways , and to keep his commandments , and to cleave unto him , and to serve him

to 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 .

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

16 : 14 and the pharisees

uppermost

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

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

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

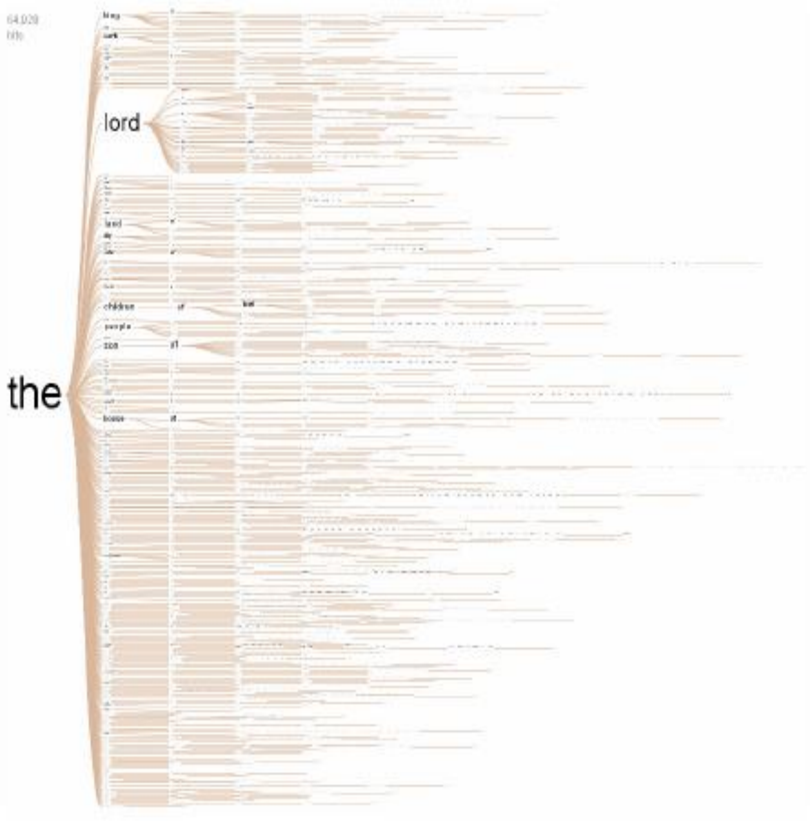
brotherhood .

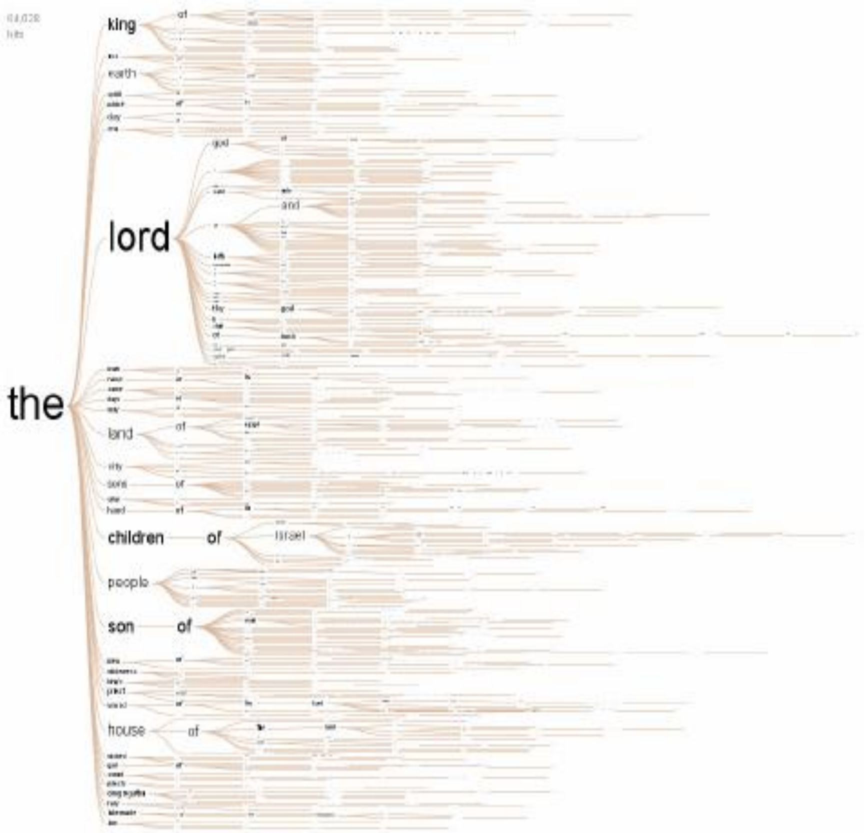
world , the love of the father is not in him .

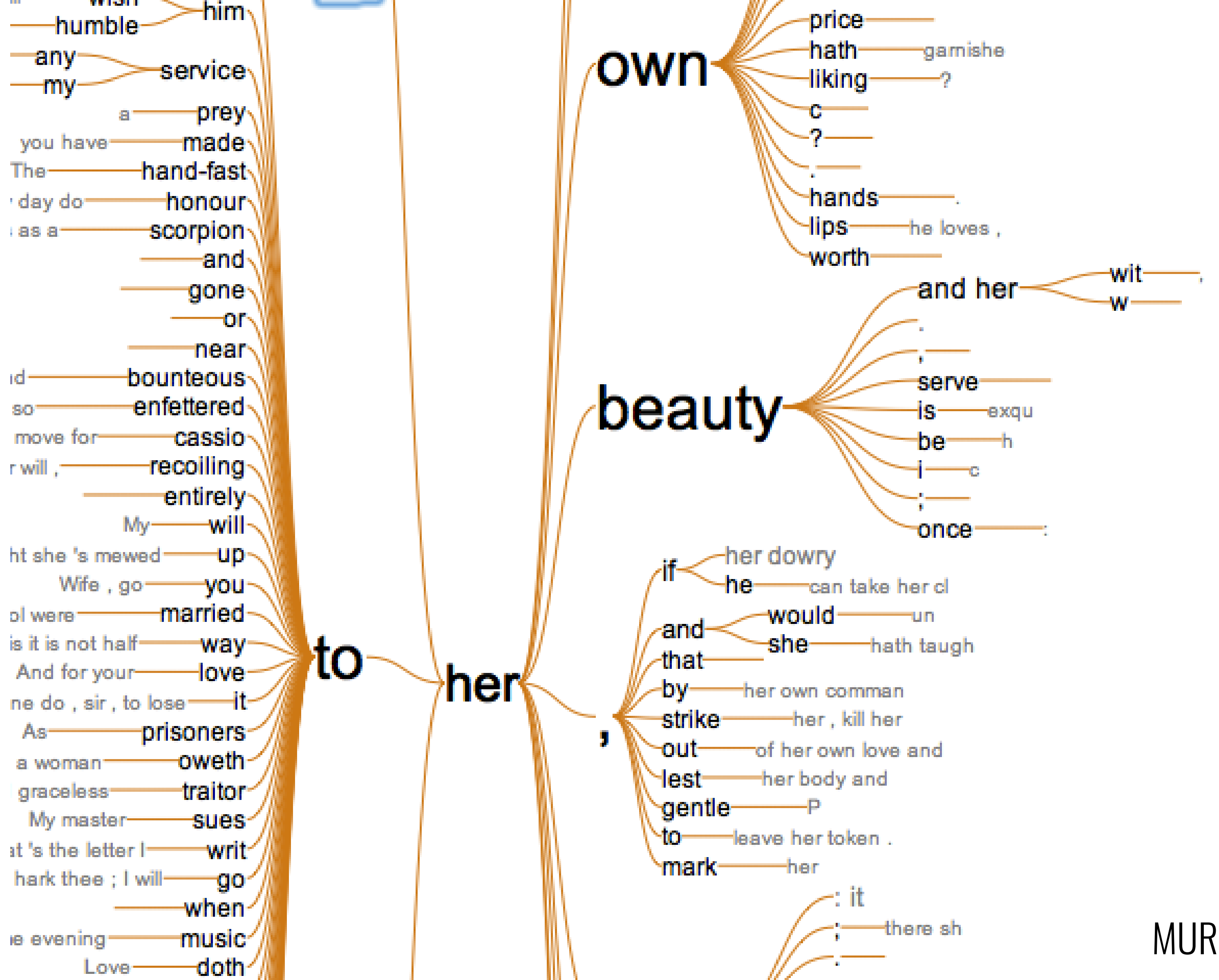
brethren .

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

FILTER INFREQUENT RUNS

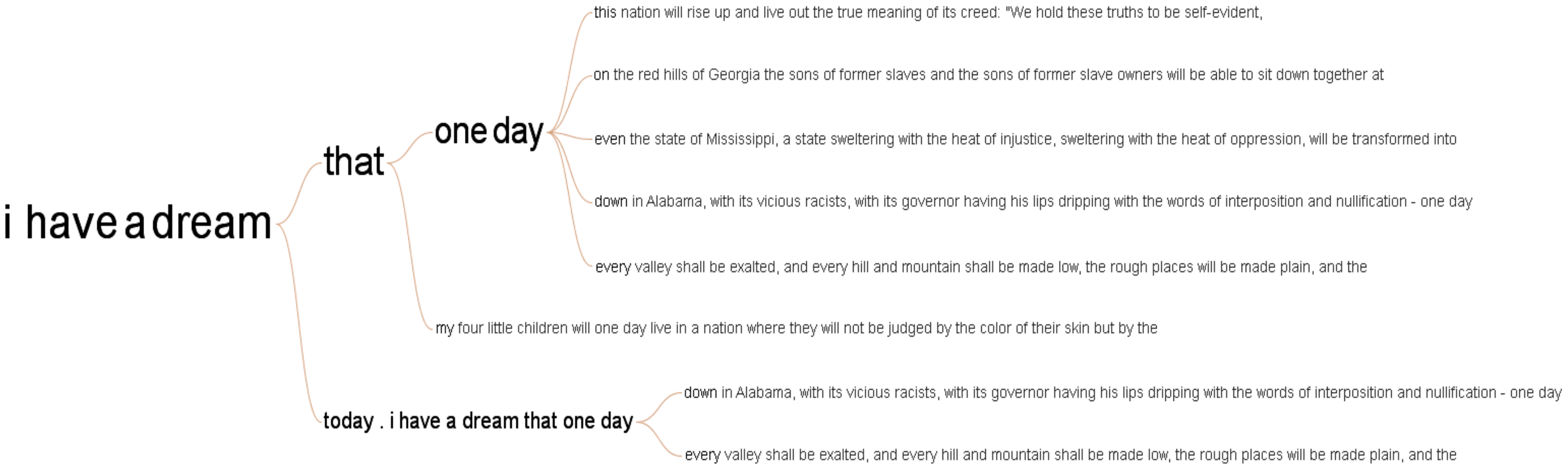






WORDSEER
MURALIDHARAN & HEARST

RECURRENT THEMES IN SPEECH



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

GLIMPSES OF STRUCTURE

- CONCORDANCES SHOW LOCAL, REPEATED STRUCTURE
- BUT WHAT ABOUT OTHER TYPES OF PATTERNS?

- FOR EXAMPLE
- LEXICAL: <A> at <B>
- SYNTACTIC: <Noun> <Verb> <Object>

PHRASE NETS

LOOK FOR SPECIFIC LINKING PATTERNS IN THE TEXT:

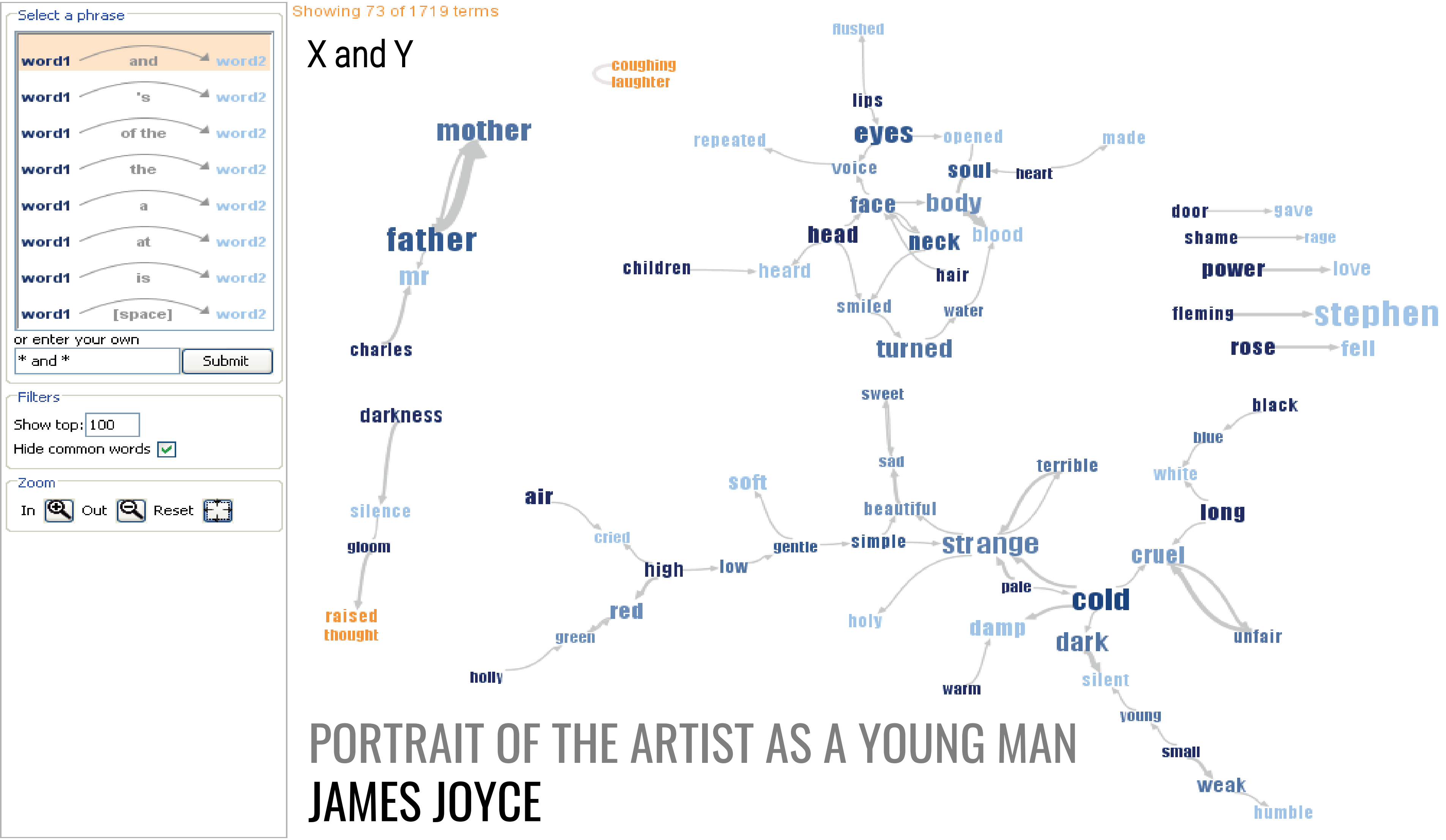
- ‘A AND B’, ‘A AT B’, ‘A OF B’, ETC
- COULD BE OUTPUT OF REGEXP OR PARSER

VISUALIZE EXTRACTED PATTERNS IN A NODE-LINK VIEW

- OCCURRENCES = NODE SIZE
- PATTERN POSITION = EDGE DIRECTION

van Ham et al

X and Y



Showing 73 of 1719 terms

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

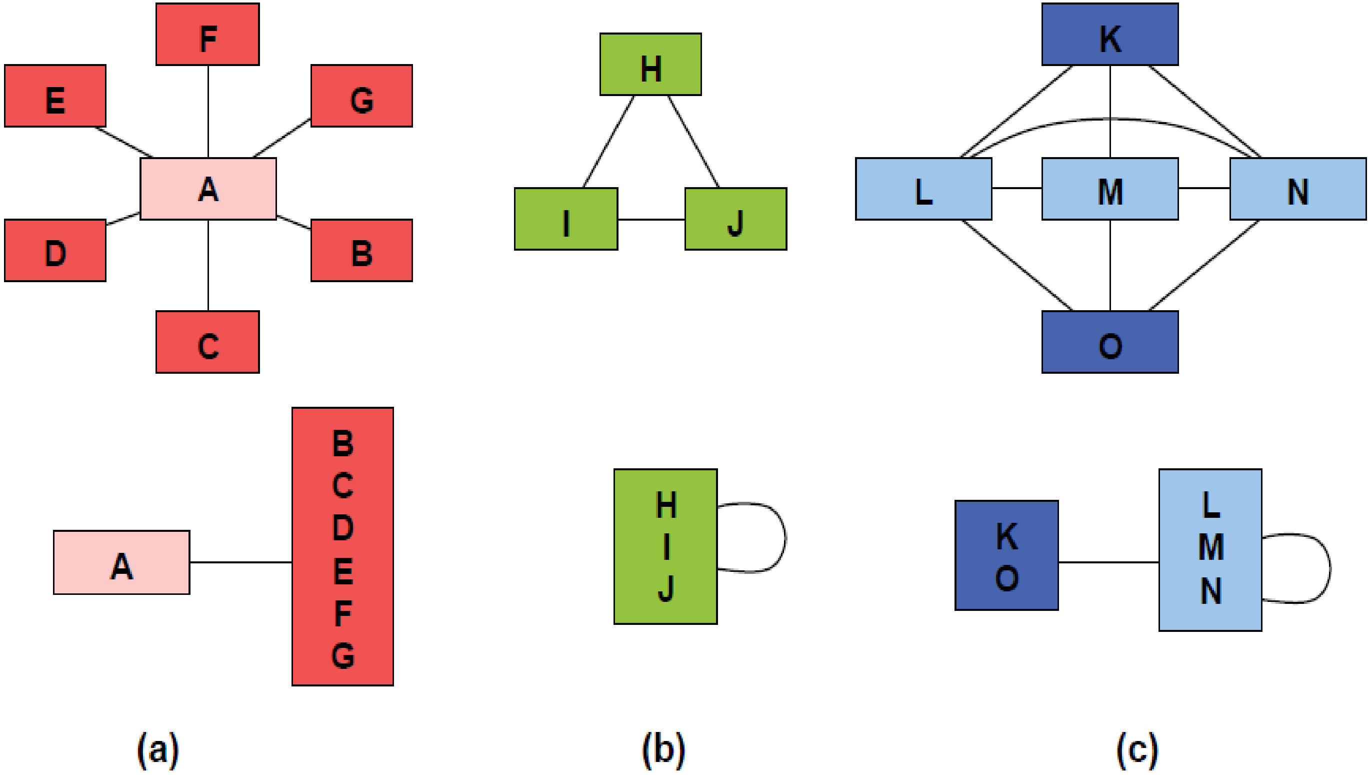
or enter your own
* and *
Submit

Filters
Show top: 100
Hide common words ☑

Zoom
In Out Reset

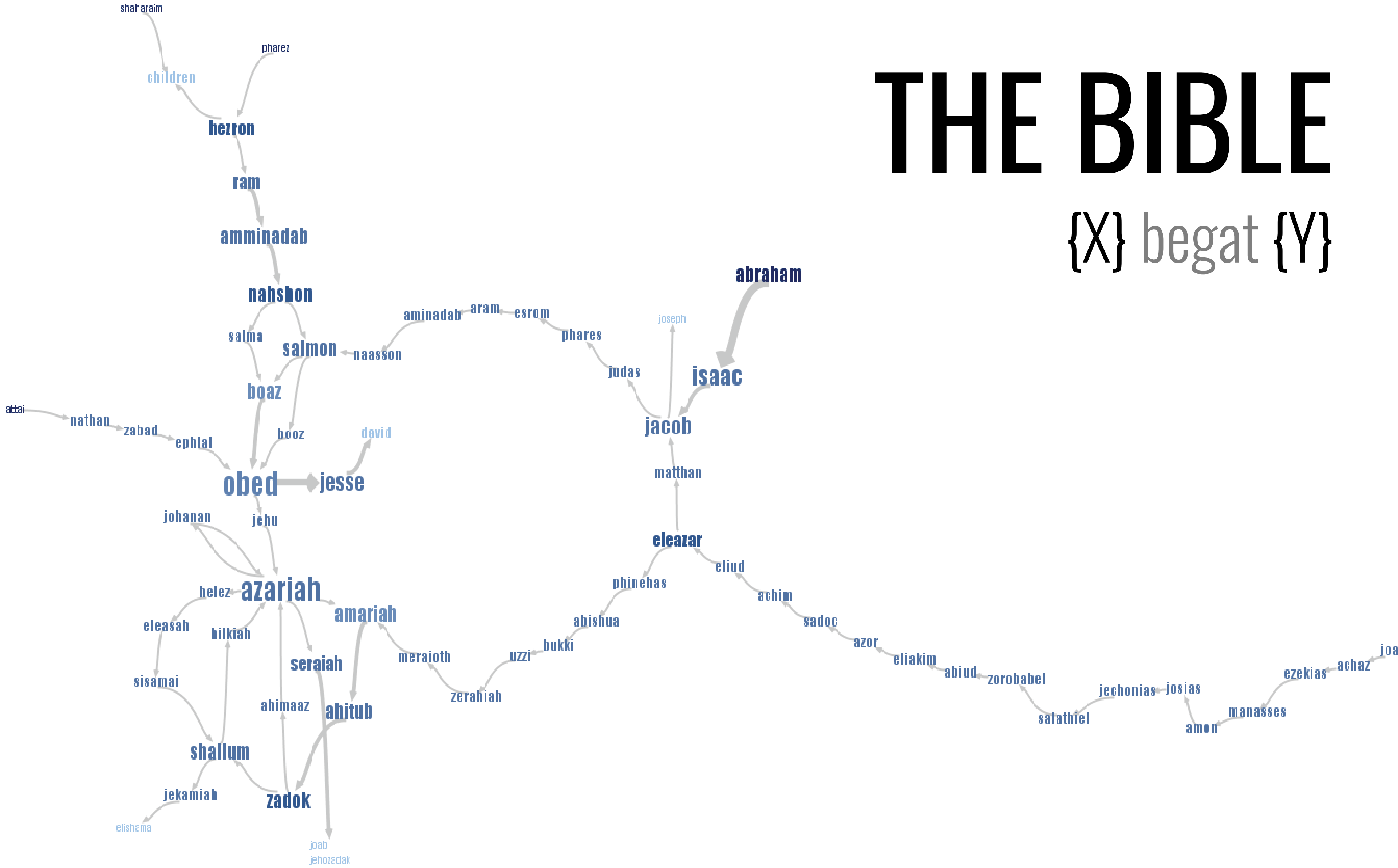
PORTRAIT OF THE ARTIST AS A YOUNG MAN
JAMES JOYCE

NODE GROUPING



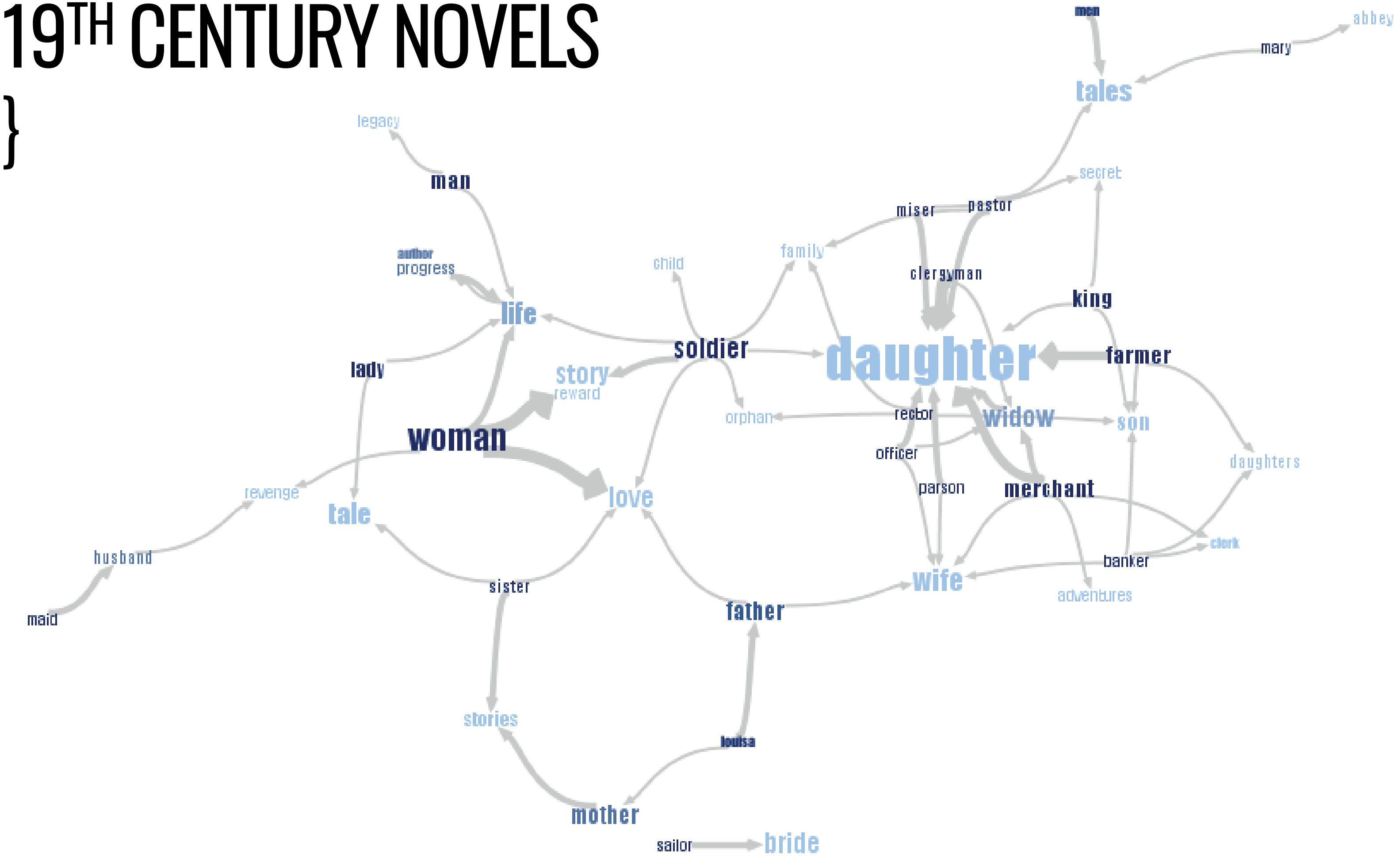
THE BIBLE

{X} begat {Y}



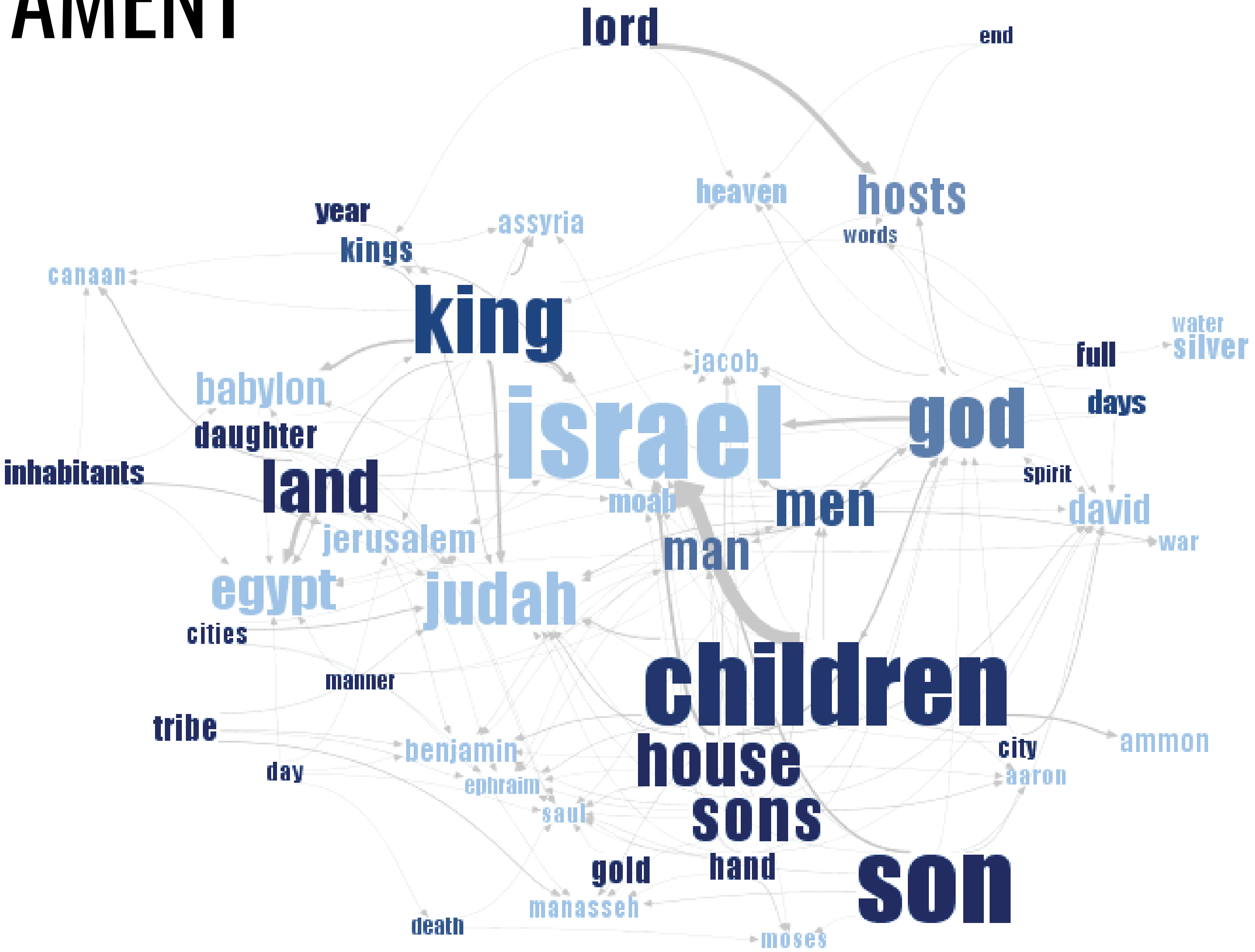
18TH & 19TH CENTURY NOVELS

{X}'s {Y}



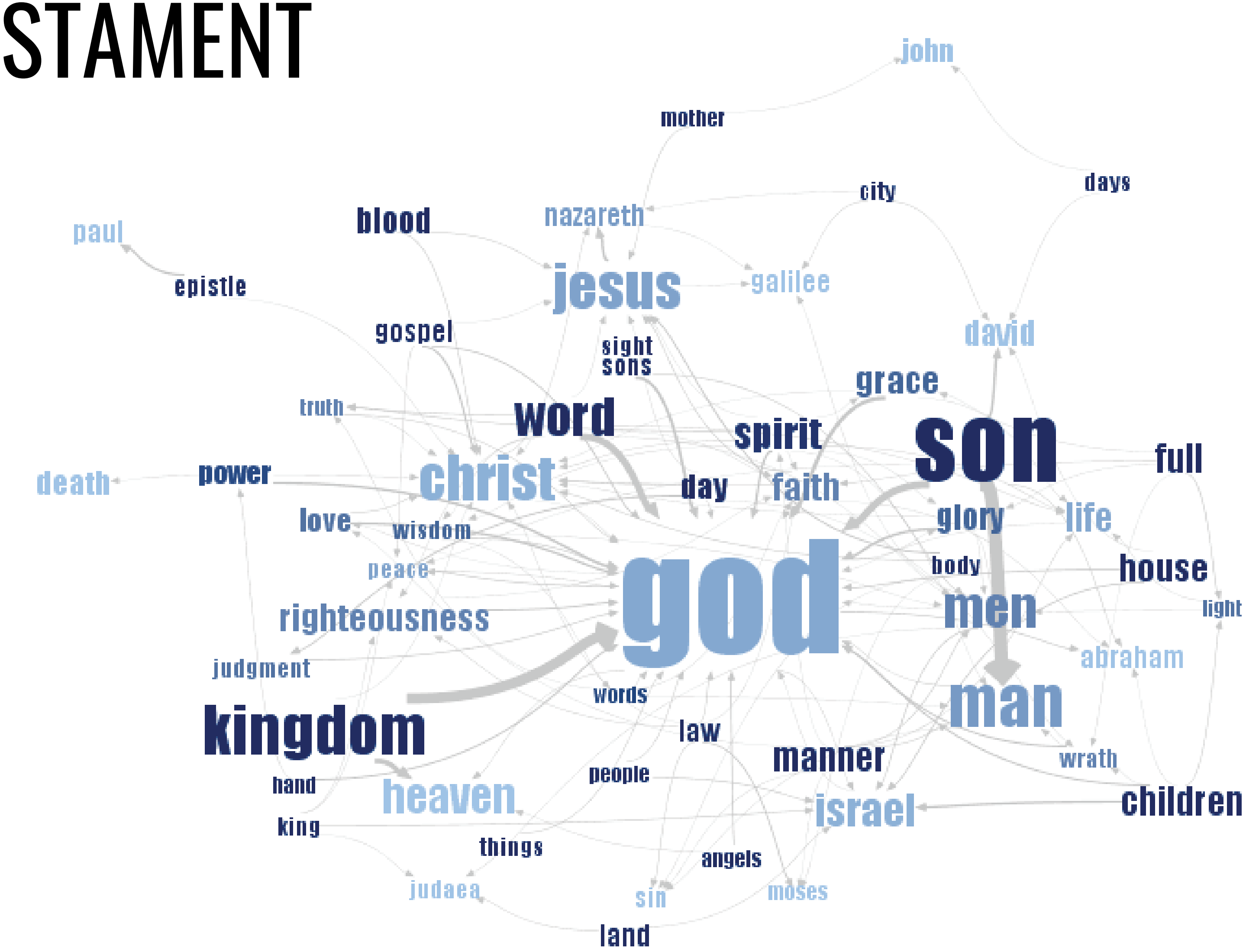
OLD TESTAMENT

{X} of {Y}



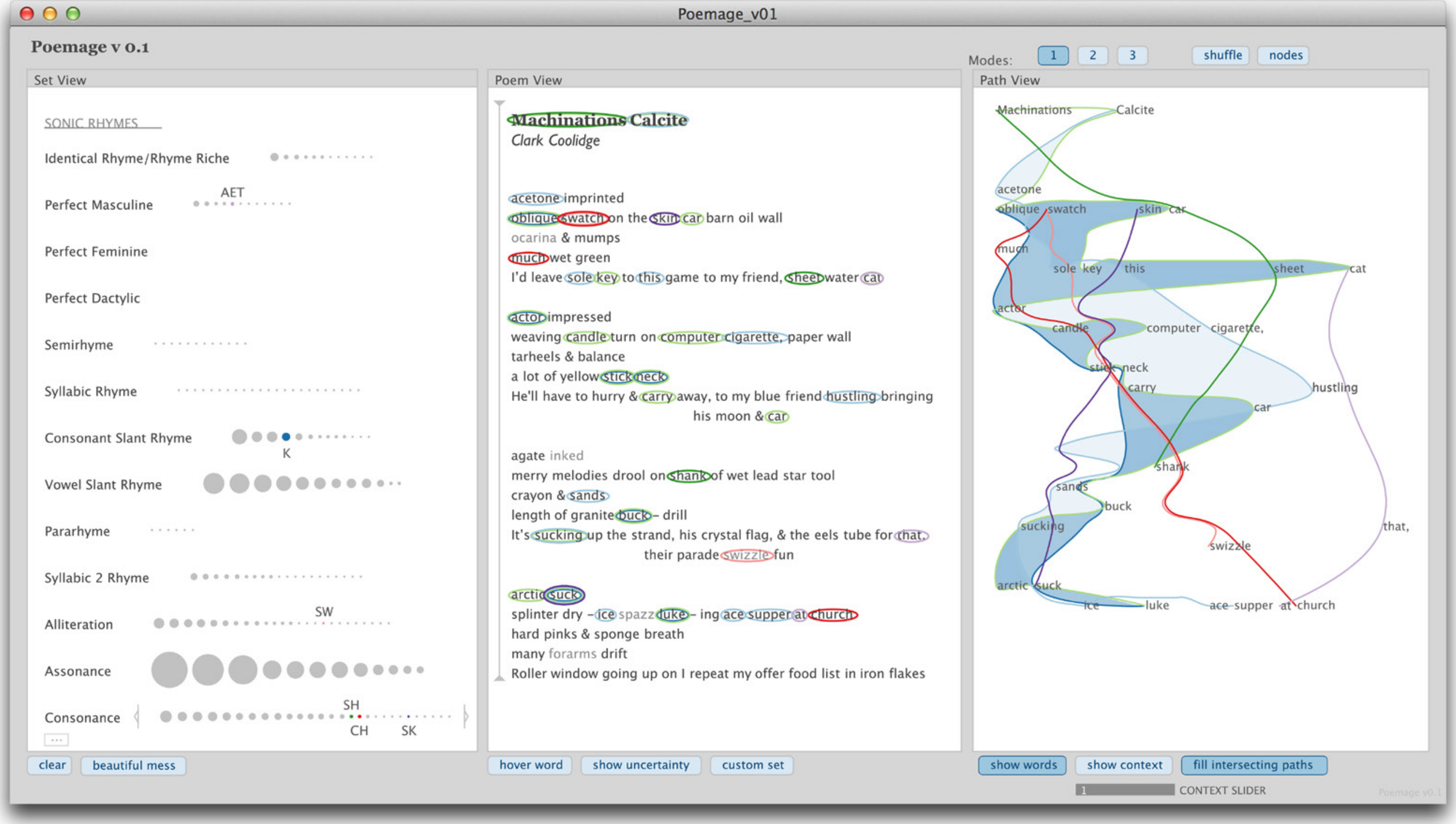
NEW TESTAMENT

{X} of {Y}



RHYME, SPEECH, ETC.

POEMAGE McCurdy et al. 2016



Machinations Calcite

Clark Coolidge

acetone imprinted
oblique swatch on the skin car barn oil wall
ocarina & mumps
much wet green
I'd leave sole key to this game to my friend, sheet water cat

actor impressed
weaving candle turn on computer cigarette, paper wall
tarheels & balance
a lot of yellow stick neck
He'll have to hurry & carry away, to my blue friend hustling bringing
his moon & car

agate inked
merry melodies drool on shank of wet lead star tool
crayon & sands
length of granite buck – drill
It's sucking up the strand, his crystal flag, & the eels tube for that,
their parade swizzle fun

arctic suck
splinter dry – ice spazz luke – ing ace supper at church
hard pinks & sponge breath
many forarms drift
Roller window going up on I repeat my offer food list in iron flakes

REVISIT YOUR SKETCHES?

TASK:

1) VISUALIZE THE MOST IMPORTANT CONTENT FROM A SINGLE THESIS.

ARE YOUR VISUALIZATION CHOICES **EFFECTIVE**?

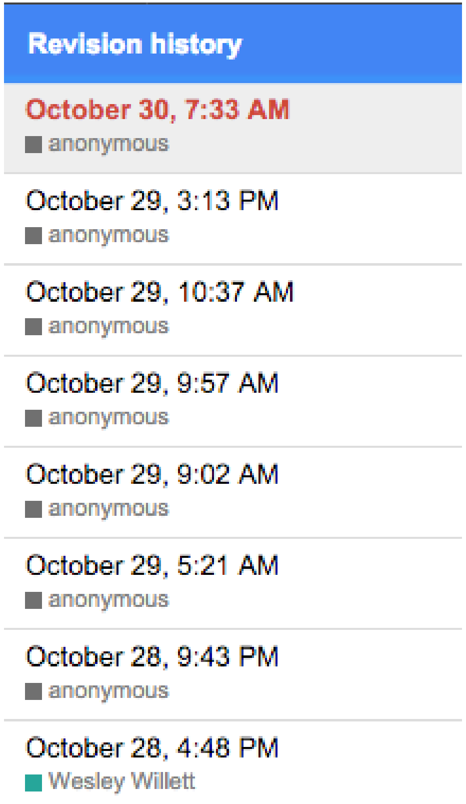
DOES THE VIS CAPTURE THE **LENGTH**, **FORM**, AND **POSITION** OF THE IMPORTANT CONTENT?

DO YOU SHOW OR CONNECT BACK TO THE **ORIGINAL TEXT**?

EVOLVING DOCUMENTS

VISUALIZING REVISION HISTORY

HOW TO DEPICT CONTRIBUTIONS AND CHANGES OVER TIME?



Revision history

October 30, 7:33 AM
▇ anonymous

October 29, 3:13 PM
▇ anonymous

October 29, 10:37 AM
▇ anonymous

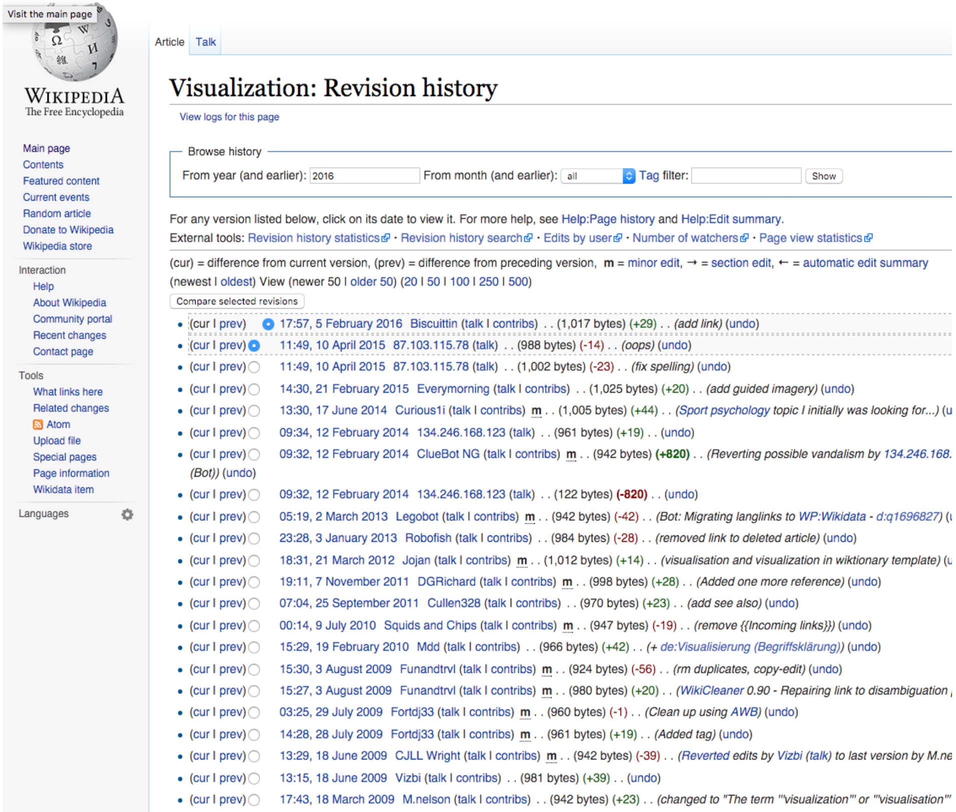
October 29, 9:57 AM
▇ anonymous

October 29, 9:02 AM
▇ anonymous

October 29, 5:21 AM
▇ anonymous

October 28, 9:43 PM
▇ anonymous

October 28, 4:48 PM
🟩 Wesley Willett



Visualization: Revision history

View logs for this page

Browse history

From year (and earlier): 2016 From month (and earlier): all Tag filter: Show

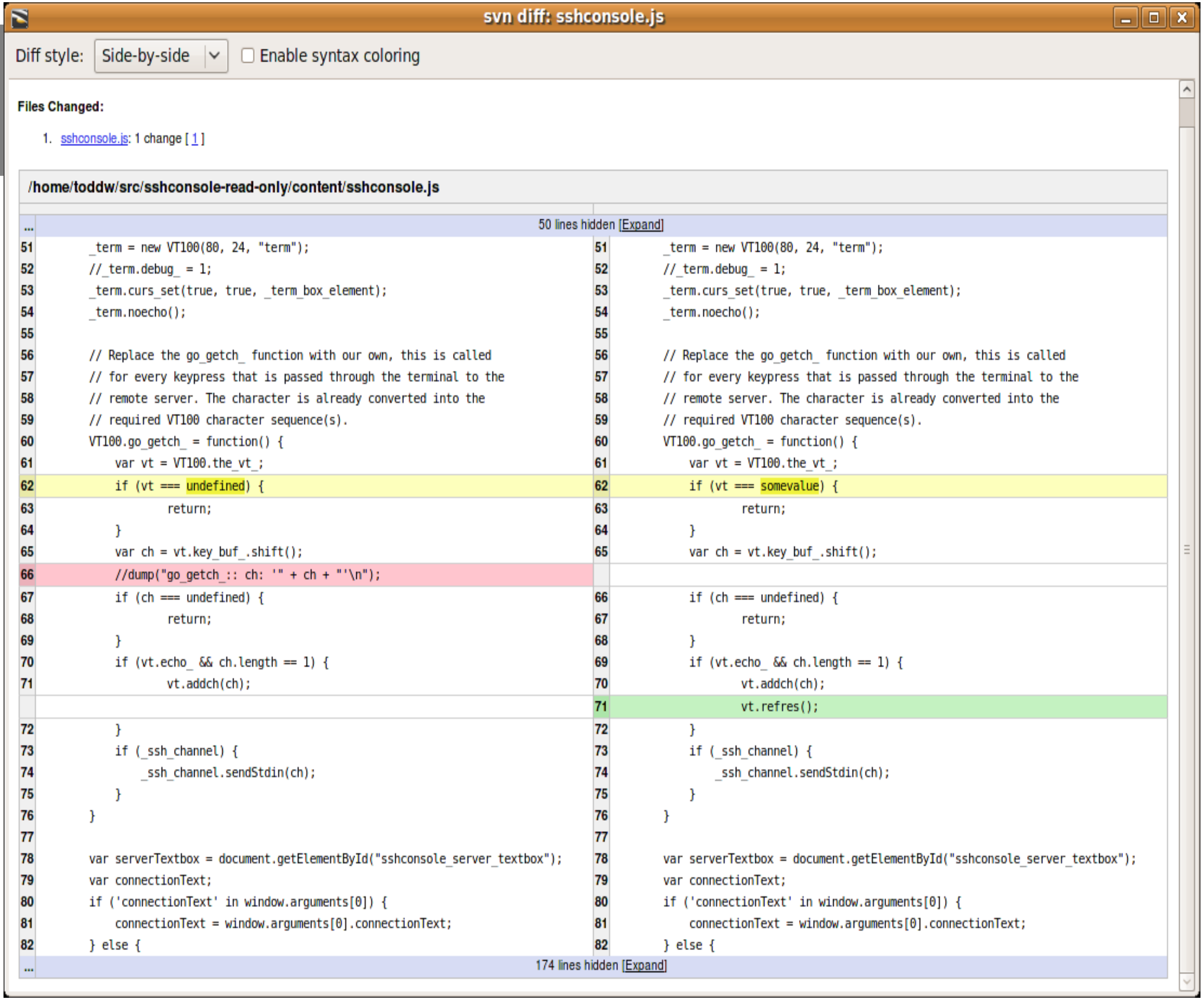
For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary.
External tools: Revision history statistics · Revision history search · Edits by user · Number of watchers · Page view statistics

(cur) = difference from current version, (prev) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary
(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur | prev) ● 17:57, 5 February 2016 Biscuittin (talk | contribs) . . (1,017 bytes) (+29) . . *(add link)* (undo)
- (cur | prev) ● 11:49, 10 April 2015 87.103.115.78 (talk) . . (988 bytes) (-14) . . *(oops)* (undo)
- (cur | prev) ○ 11:49, 10 April 2015 87.103.115.78 (talk) . . (1,002 bytes) (-23) . . *(fix spelling)* (undo)
- (cur | prev) ○ 14:30, 21 February 2015 Everymorning (talk | contribs) . . (1,025 bytes) (+20) . . *(add guided imagery)* (undo)
- (cur | prev) ○ 13:30, 17 June 2014 Curious1i (talk | contribs) **m** . . (1,005 bytes) (+44) . . *(Sport psychology topic I initially was looking for...)* (u
- (cur | prev) ○ 09:34, 12 February 2014 134.246.168.123 (talk) . . (961 bytes) (+19) . . (undo)
- (cur | prev) ○ 09:32, 12 February 2014 ClueBot NG (talk | contribs) **m** . . (942 bytes) **(+820)** . . *(Reverting possible vandalism by 134.246.168. (Bot))* (undo)
- (cur | prev) ○ 09:32, 12 February 2014 134.246.168.123 (talk) . . (122 bytes) **(-820)** . . (undo)
- (cur | prev) ○ 05:19, 2 March 2013 Legobot (talk | contribs) **m** . . (942 bytes) (-42) . . *(Bot: Migrating langlinks to WP:Wikidata - d:q1696827)* (
- (cur | prev) ○ 23:28, 3 January 2013 Robofish (talk | contribs) . . (984 bytes) (-28) . . *(removed link to deleted article)* (undo)
- (cur | prev) ○ 18:31, 21 March 2012 Jojan (talk | contribs) **m** . . (1,012 bytes) (+14) . . *(visualisation and visualization in wiktionary template)* (u
- (cur | prev) ○ 19:11, 7 November 2011 DGRichard (talk | contribs) **m** . . (998 bytes) (+28) . . *(Added one more reference)* (undo)
- (cur | prev) ○ 07:04, 25 September 2011 Cullen328 (talk | contribs) . . (970 bytes) (+23) . . *(add see also)* (undo)
- (cur | prev) ○ 00:14, 9 July 2010 Squids and Chips (talk | contribs) **m** . . (947 bytes) (-19) . . *(remove {{Incoming links}})* (undo)
- (cur | prev) ○ 15:29, 19 February 2010 Mdd (talk | contribs) . . (966 bytes) (+42) . . *(+ de:Visualisierung (Begriffsklärung))* (undo)
- (cur | prev) ○ 15:30, 3 August 2009 Funandtrvl (talk | contribs) **m** . . (924 bytes) (-56) . . *(rm duplicates, copy-edit)* (undo)
- (cur | prev) ○ 15:27, 3 August 2009 Funandtrvl (talk | contribs) **m** . . (980 bytes) (+20) . . *(WikiCleaner 0.90 - Repairing link to disambiguation*
- (cur | prev) ○ 03:25, 29 July 2009 Fortdj33 (talk | contribs) **m** . . (960 bytes) (-1) . . *(Clean up using AWB)* (undo)
- (cur | prev) ○ 14:28, 28 July 2009 Fortdj33 (talk | contribs) **m** . . (961 bytes) (+19) . . *(Added tag)* (undo)
- (cur | prev) ○ 13:29, 18 June 2009 CJLL Wright (talk | contribs) **m** . . (942 bytes) (-39) . . *(Reverted edits by Vizbi (talk) to last version by M.ne*
- (cur | prev) ○ 13:15, 18 June 2009 Vizbi (talk | contribs) . . (981 bytes) (+39) . . (undo)
- (cur | prev) ○ 17:43, 18 March 2009 M.nelson (talk | contribs) . . (942 bytes) (+23) . . *(changed to "The term '''visualization''' or '''visualisation'''*

DIFF



Files Changed:

1. sshconsole.js: 1 change [1]

/home/toddw/src/sshconsole-read-only/content/sshconsole.js

50 lines hidden [Expand]

```
        _term = new VT100(80, 24, "term");
        //_term.debug_ = 1;
        _term.curs_set(true, true, _term_box_element);
        _term.noecho();

        // Replace the go_getch_ function with our own, this is called
        // for every keypress that is passed through the terminal to the
        // remote server. The character is already converted into the
        // required VT100 character sequence(s).
        VT100.go_getch_ = function() {
            var vt = VT100.the_vt_;
            if (vt === undefined) {
                    return;
            }
            var ch = vt.key_buf_.shift();
            //dump("go_getch_:: ch: '" + ch + "'\n");
            if (ch === undefined) {
                    return;
            }
            if (vt.echo_ && ch.length == 1) {
                    vt.addch(ch);
            }
            if (_ssh_channel) {
                _ssh_channel.sendStdin(ch);
            }
        }

        var serverTextbox = document.getElementById("sshconsole_server_textbox");
        var connectionText;
        if ('connectionText' in window.arguments[0]) {
            connectionText = window.arguments[0].connectionText;
        } else {
```

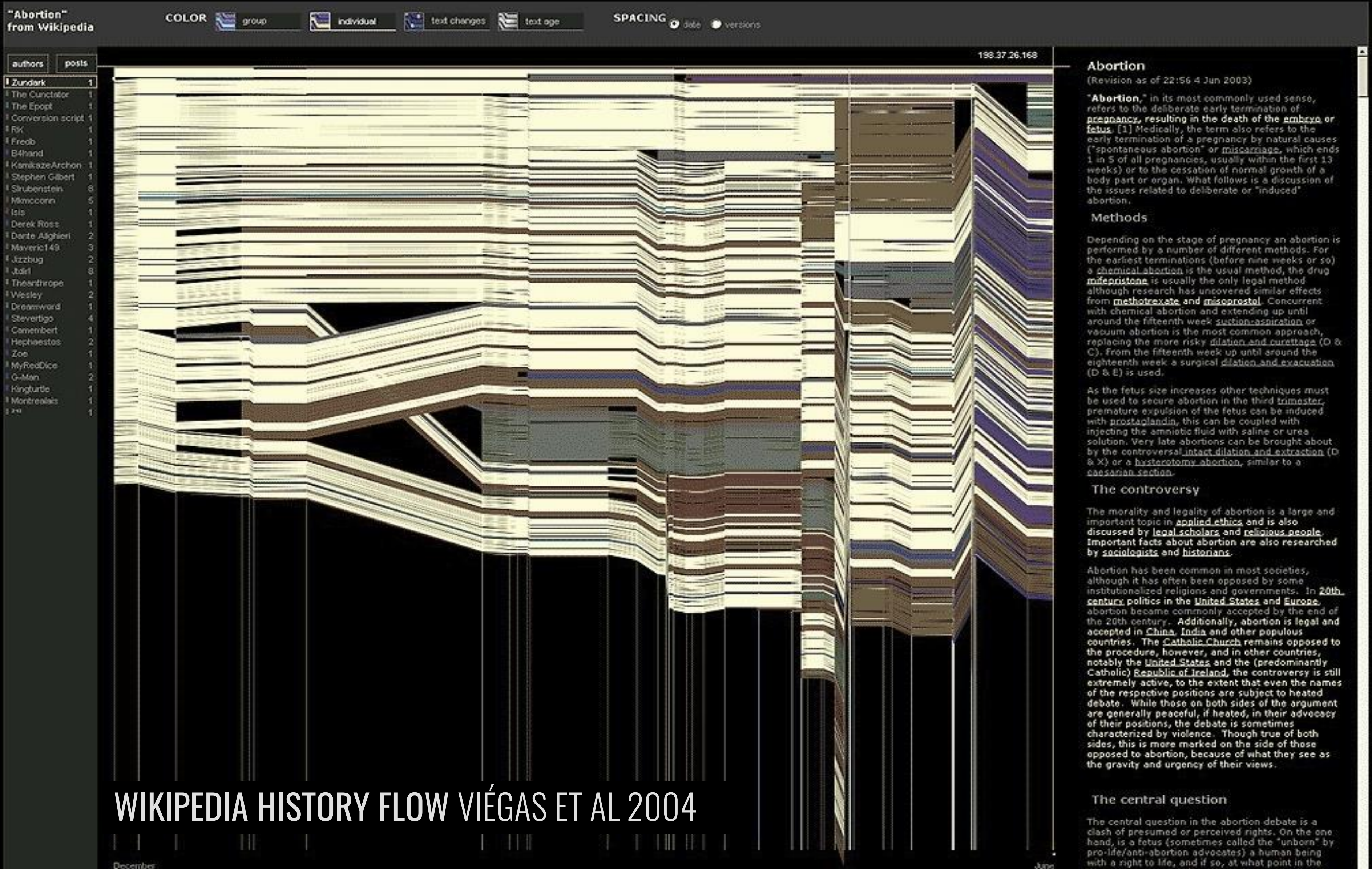
```
        _term = new VT100(80, 24, "term");
        //_term.debug_ = 1;
        _term.curs_set(true, true, _term_box_element);
        _term.noecho();

        // Replace the go_getch_ function with our own, this is called
        // for every keypress that is passed through the terminal to the
        // remote server. The character is already converted into the
        // required VT100 character sequence(s).
        VT100.go_getch_ = function() {
            var vt = VT100.the_vt_;
            if (vt === somevalue) {
                    return;
            }
            var ch = vt.key_buf_.shift();

            if (ch === undefined) {
                    return;
            }
            if (vt.echo_ && ch.length == 1) {
                    vt.addch(ch);
                    vt.refres();
            }
            if (_ssh_channel) {
                _ssh_channel.sendStdin(ch);
            }
        }

        var serverTextbox = document.getElementById("sshconsole_server_textbox");
        var connectionText;
        if ('connectionText' in window.arguments[0]) {
            connectionText = window.arguments[0].connectionText;
        } else {
```

174 lines hidden [Expand]



Zundark	1
The Cunctator	1
The Epopt	1
Conversion script	1
RK	1
Fredb	1
B4hand	1
KamikazeArchon	1
Stephen Gilbert	1
Slrubenstein	8
Mkmcconn	5
Isis	1
Derek Ross	1
Dante Alighieri	2
Maveric149	3
Jizzbug	2
Jtdirl	8
Theanthrope	1
Wesley	2
Dreamword	1
Stevertigo	4
Camembert	1
Hephaestos	2
Zoe	1
MyRedDice	1
G-Man	2
Kingturtle	1
Montrealais	1
2·0	1

Abortion

(Revision as of 22:56 4 Jun 2003)

"**Abortion**," in its most commonly used sense, refers to the deliberate early termination of pregnancy, resulting in the death of the embryo or fetus. [1] Medically, the term also refers to the early termination of a pregnancy by natural causes ("spontaneous abortion" or miscarriage, which ends 1 in 5 of all pregnancies, usually within the first 13 weeks) or to the cessation of normal growth of a body part or organ. What follows is a discussion of the issues related to deliberate or "induced" abortion.

Methods

Depending on the stage of pregnancy an abortion is performed by a number of different methods. For the earliest terminations (before nine weeks or so) a chemical abortion is the usual method, the drug mifepristone is usually the only legal method although research has uncovered similar effects from methotrexate and misoprostol. Concurrent with chemical abortion and extending up until around the fifteenth week suction-aspiration or vacuum abortion is the most common approach, replacing the more risky dilation and curettage (D & C). From the fifteenth week up until around the eighteenth week a surgical dilation and evacuation (D & E) is used.

As the fetus size increases other techniques must be used to secure abortion in the third trimester, premature expulsion of the fetus can be induced with prostaglandin, this can be coupled with injecting the amniotic fluid with saline or urea solution. Very late abortions can be brought about by the controversal intact dilation and extraction (D & X) or a hysterotomy abortion, similar to a caesarian section.

The controversy

The morality and legality of abortion is a large and important topic in applied ethics and is also discussed by legal scholars and religious people. Important facts about abortion are also researched by sociologists and historians.

Abortion has been common in most societies, although it has often been opposed by some institutionalized religions and governments. In 20th century politics in the United States and Europe, abortion became commonly accepted by the end of the 20th century. Additionally, abortion is legal and accepted in China, India and other populous countries. The Catholic Church remains opposed to the procedure, however, and in other countries, notably the United States and the (predominantly Catholic) Republic of Ireland, the controversy is still extremely active, to the extent that even the names of the respective positions are subject to heated debate. While those on both sides of the argument are generally peaceful, if heated, in their advocacy of their positions, the debate is sometimes characterized by violence. Though true of both sides, this is more marked on the side of those opposed to abortion, because of what they see as the gravity and urgency of their views.

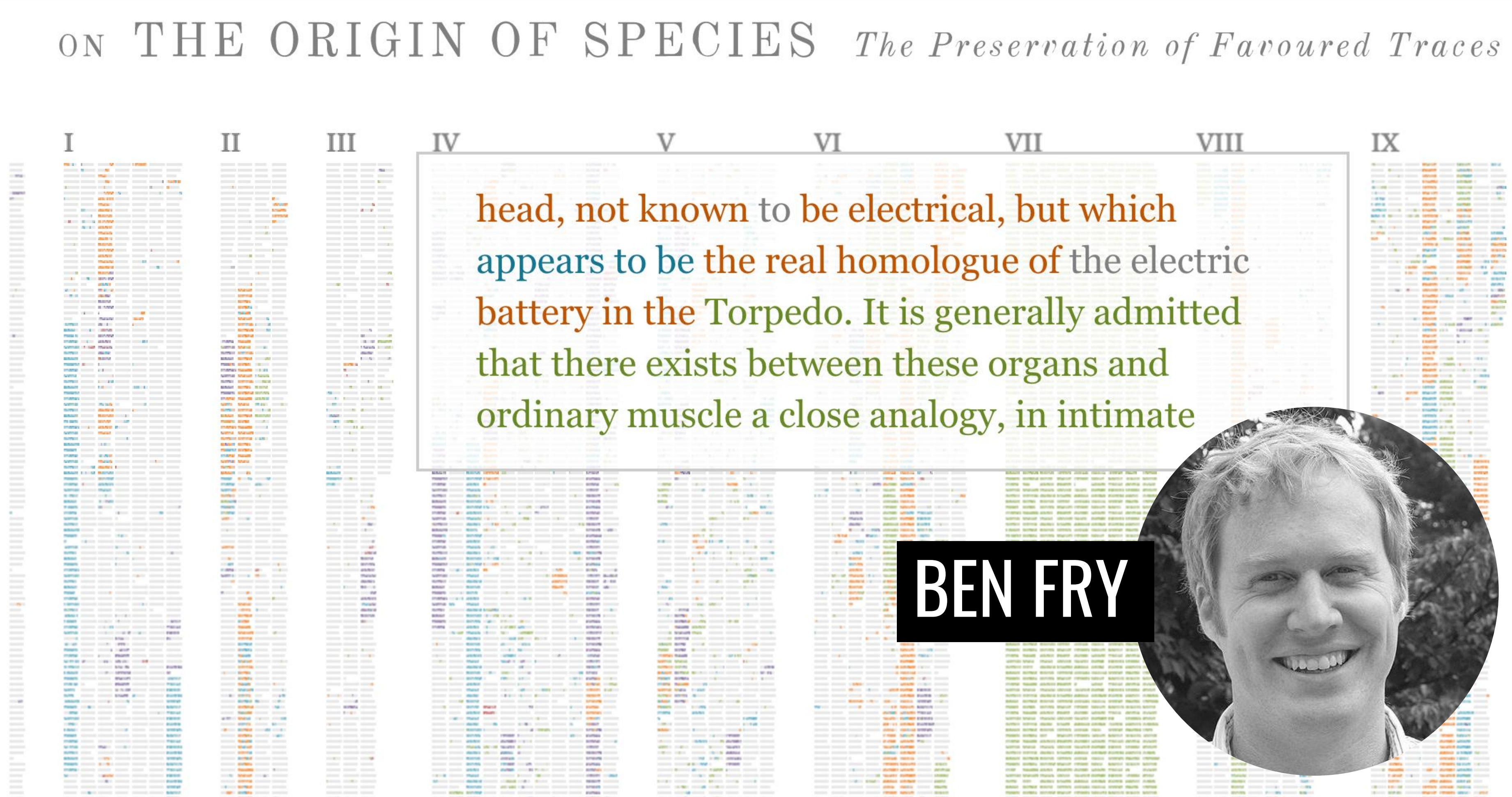
The central question

The central question in the abortion debate is a clash of presumed or perceived rights. On the one hand, is a fetus (sometimes called the "unborn" by pro-life/anti-abortion advocates) a human being with a right to life, and if so, at what point in the

WIKIPEDIA HISTORY FLOW VIÉGAS ET AL 2004

ANIMATED TRACES

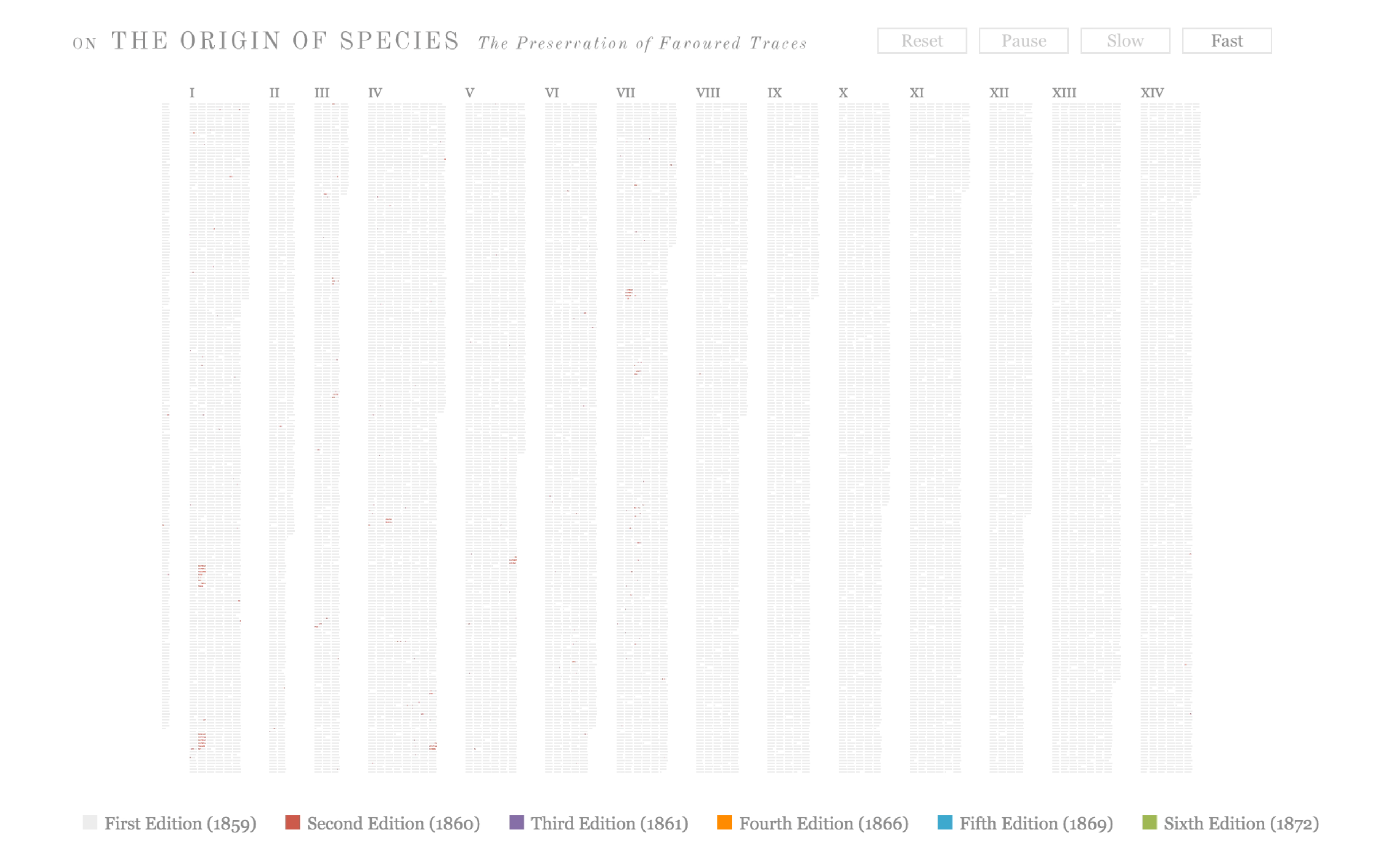
fathom.info/traces



ON THE ORIGIN OF SPECIES *The Preservation of Favoured Traces*

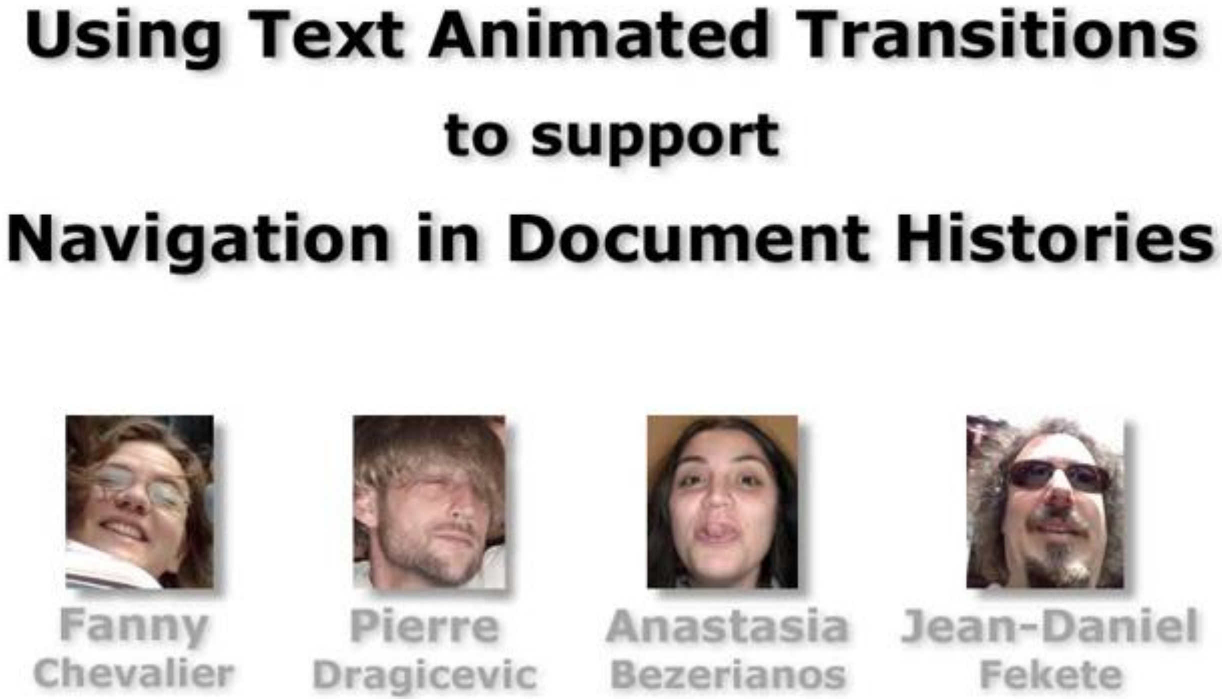
head, not known to be electrical, but which appears to be the real homologue of the electric battery in the Torpedo. It is generally admitted that there exists between these organs and ordinary muscle a close analogy, in intimate

BEN FRY



ON THE ORIGIN OF SPECIES *The Preservation of Favoured Traces*

DIFFAMATION



Using Text Animated Transitions

to support

Navigation in Document Histories

Fanny Chevalier

Pierre Dragicevic

Anastasia Bezerianos

Jean-Daniel Fekete

Shortest Edit Path

Edit War

VISUALIZING DOCUMENT COLLECTIONS

SKETCHING: **VISUALIZE**

IMAGINE YOU HAVE A MASTER’S THESIS IN FROM OF YOU:

YEAR

AUTHOR

TITLE

KEYWORDS

REFERENCES

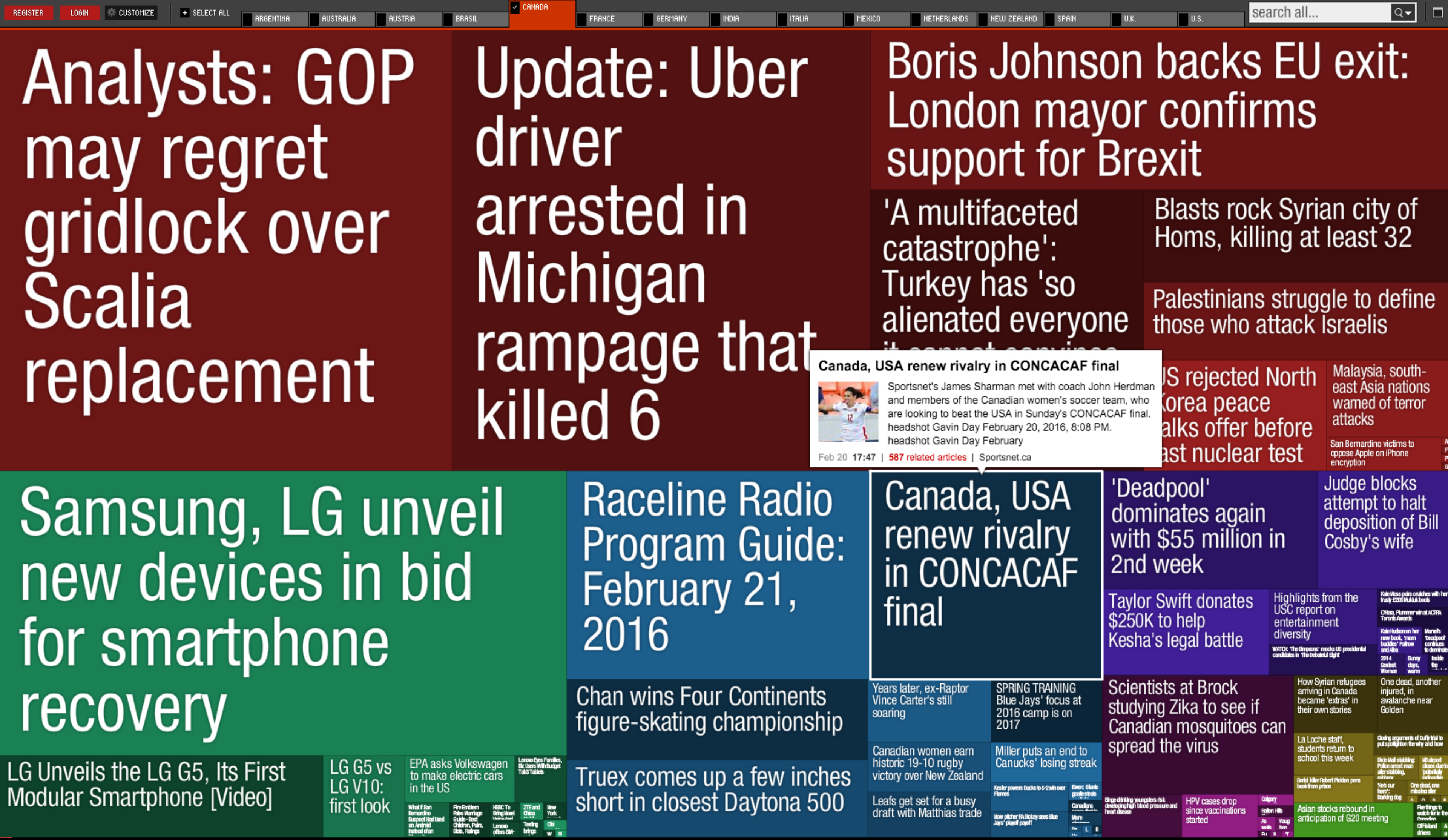
ABSTRACT TEXT

TASK:

1) VISUALIZE THE MOST IMPORTANT CONTENT FROM A SINGLE THESIS.

2) VISUALIZE HOW SIMILAR THIS THESIS IS TO THESES FROM OTHER STUDENTS IN THE SAME MASTER’S PROGRAM.

(~10 MINUTES)



Analysts: GOP may regret gridlock over Scalia replacement

Update: Uber driver arrested in Michigan rampage that killed 6

Boris Johnson backs EU exit: London mayor confirms support for Brexit

'A multifaceted catastrophe': Turkey has 'so alienated everyone it cannot survive

Blasts rock Syrian city of Homs, killing at least 32

Palestinians struggle to define those who attack Israelis

US rejected North Korea peace talks offer before last nuclear test

Malaysia, south-east Asia nations warned of terror attacks

San Bernardino victims to oppose Apple on iPhone encryption

Canada, USA renew rivalry in CONCACAF final

Sportsnet's James Sharman met with coach John Herdman and members of the Canadian women's soccer team, who are looking to beat the USA in Sunday's CONCACAF final. headshot Gavin Day February 20, 2016, 8:08 PM. headshot Gavin Day February

Feb 20 17:47 | 587 related articles | Sportsnet.ca

Samsung, LG unveil new devices in bid for smartphone recovery

LG Unveils the LG G5, Its First Modular Smartphone [Video]

LG G5 vs LG V10: first look

EPA asks Volkswagen to make electric cars in the US

Lenovo Eyes Families, Biz Users With Budget Tab3 Tablets

What if San Bernardino Suspect Had Used an Android Instead of an

Fire Emblem Fates Marriage Guide - Best Children, Pairs, Stats, Ratings

HSBC To Bring Novel

ZTE and China

New York

Lenovo offers SIM-

Texting brings

Raceline Radio Program Guide: February 21, 2016

Chan wins Four Continents figure-skating championship

Truex comes up a few inches short in closest Daytona 500

Canada, USA renew rivalry in CONCACAF final

Years later, ex-Raptor Vince Carter's still soaring

SPRING TRAINING Blue Jays' focus at 2016 camp is on 2017

Canadian women earn historic 19-10 rugby victory over New Zealand

Miller puts an end to Canucks' losing streak

Leafs get set for a busy draft with Matthias trade

Kesler powers Ducks to 5-2 win over Flames

Event: Giants goalie steals

Canadians

New pitcher RA Dickey sees Blue Jays' playoff payoff

More

'Deadpool' dominates again with $55 million in 2nd week

Judge blocks attempt to halt deposition of Bill Cosby's wife

Taylor Swift donates $250K to help Kesha's legal battle

Highlights from the USC report on entertainment diversity

WATCH: 'The Simpsons' mocks US presidential candidates in 'The Debateful Eight'

Kate Moss pairs crutches with her trusty £208 Mukluk boots

O'Hara, Plummer win at ACTRA Toronto Awards

Kate Hudson on her new book, 'mom buddies' Paltrow and Alba

Marvel's 'Deadpool' continues to dominate

2014 Sexiest Woman

Sunny days, warm

Inside the

Scientists at Brock studying Zika to see if Canadian mosquitoes can spread the virus

Binge drinking youngsters risk developing high blood pressure and heart disease

HPV cases drop since vaccinations started

Calgary

Halton Hills

How Syrian refugees arriving in Canada became 'extras' in their own stories

One dead, another injured, in avalanche near Golden

La Loche staff, students return to school this week

Closing arguments of Duffy trial to put spotlight on the why and how

Dixie Mall stabbing: Police arrest man after stabbing, robbery

Serial killer Robert Pickton pens book from prison

'He's our hero': Barking dog

One dead, one missing after

Asian stocks rebound in anticipation of G20 meeting

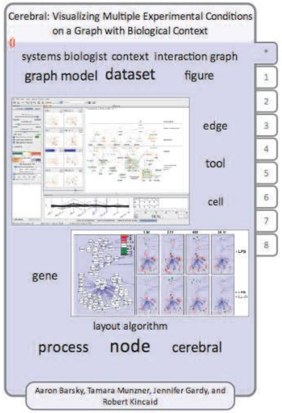
Five things to watch for in the Canadian

Off-island drivers

newsmap.jp

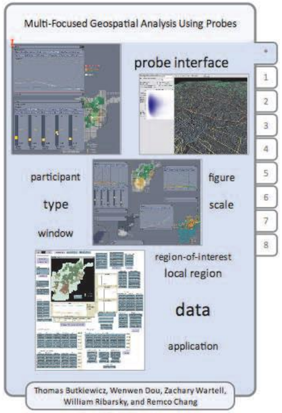
DOCUMENT CARDS

SMALL MULTIPLES FOR DOCUMENTS

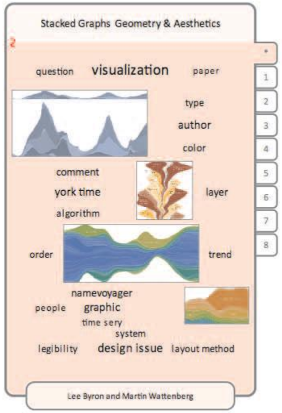

Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context

Aaron Barsky, Tamara Munzner, Jennifer Gardy, and Robert Kincaid

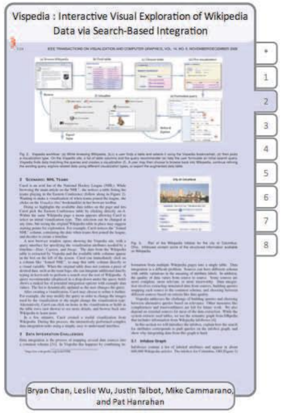
0
systems biologist context interaction graph
graph model dataset figure
edge
tool
cell
gene
layout algorithm
process node cerebral
1
2
3
4
5
6
7
8


Multi-Focused Geospatial Analysis Using Probes

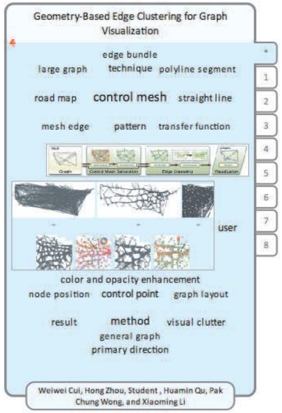
Thomas Butkiewicz, Wenwen Dou, Zachary Wartell, William Ribarsky, and Remco Chang


Stacked Graphs Geometry & Aesthetics

Lee Byron and Martin Wattenberg

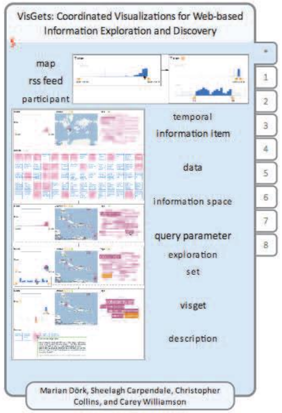

Vispedia : Interactive Visual Exploration of Wikipedia Data via Search-Based Integration

Bryan Chan, Leslie Wu, Justin Talbot, Mike Cammarano, and Pat Hanrahan

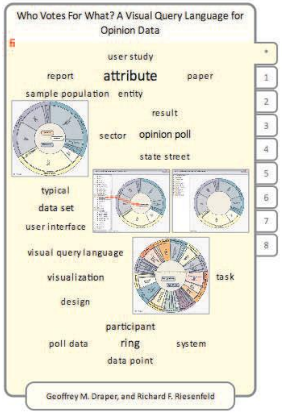

Geometry-Based Edge Clustering for Graph Visualization

Weiwei Cui, Hong Zhou, Student , Huamin Qu, Pak Chung Wong, and Xiaoming Li

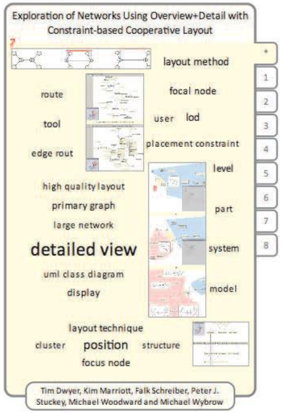
4
edge bundle
large graph technique polyline segment
road map control mesh straight line
mesh edge pattern transfer function
user
color and opacity enhancement
node position control point graph layout
result method visual clutter
general graph
primary direction


VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery

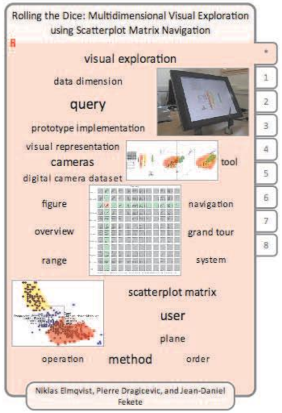
Marian Dörk, Sheelagh Carpendale, Christopher Collins, and Carey Williamson


Who Votes For What? A Visual Query Language for Opinion Data

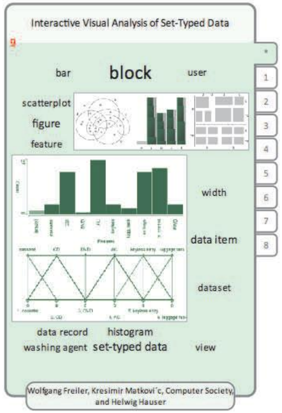
Geoffrey M. Draper, and Richard F. Riesenfeld


Exploration of Networks Using Overview+Detail with Constraint-based Cooperative Layout

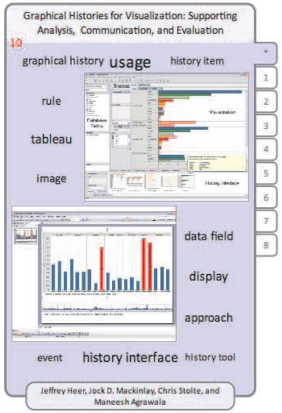
Tim Dwyer, Kim Marriott, Falk Schreiber, Peter J. Stuckey, Michael Woodward and Michael Wybrow


Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation

Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete

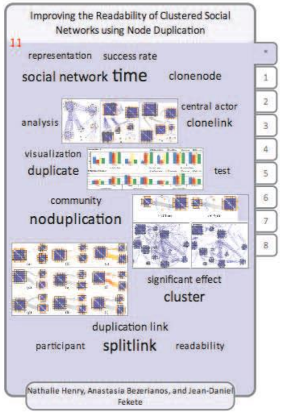

Interactive Visual Analysis of Set-Typed Data

Wolfgang Freiler, Kresimir Matkovi´c, Computer Society, and Helwig Hauser


Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation

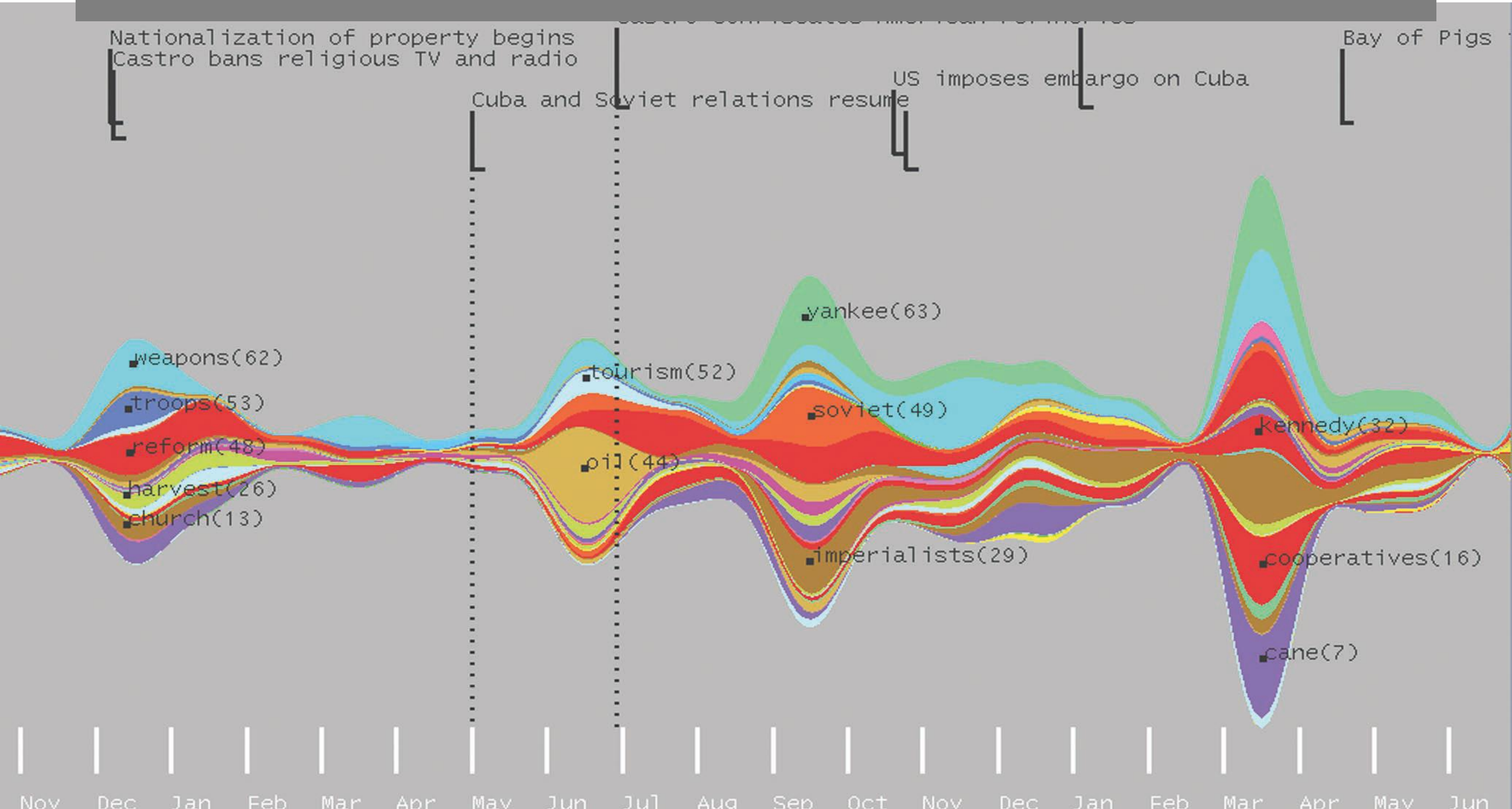
Jeffrey Heer, Jock D. Mackinlay, Chris Stolte, and Maneesh Agrawala

10
graphical history usage history item
rule
tableau
image
data field
display
approach
event history interface history tool

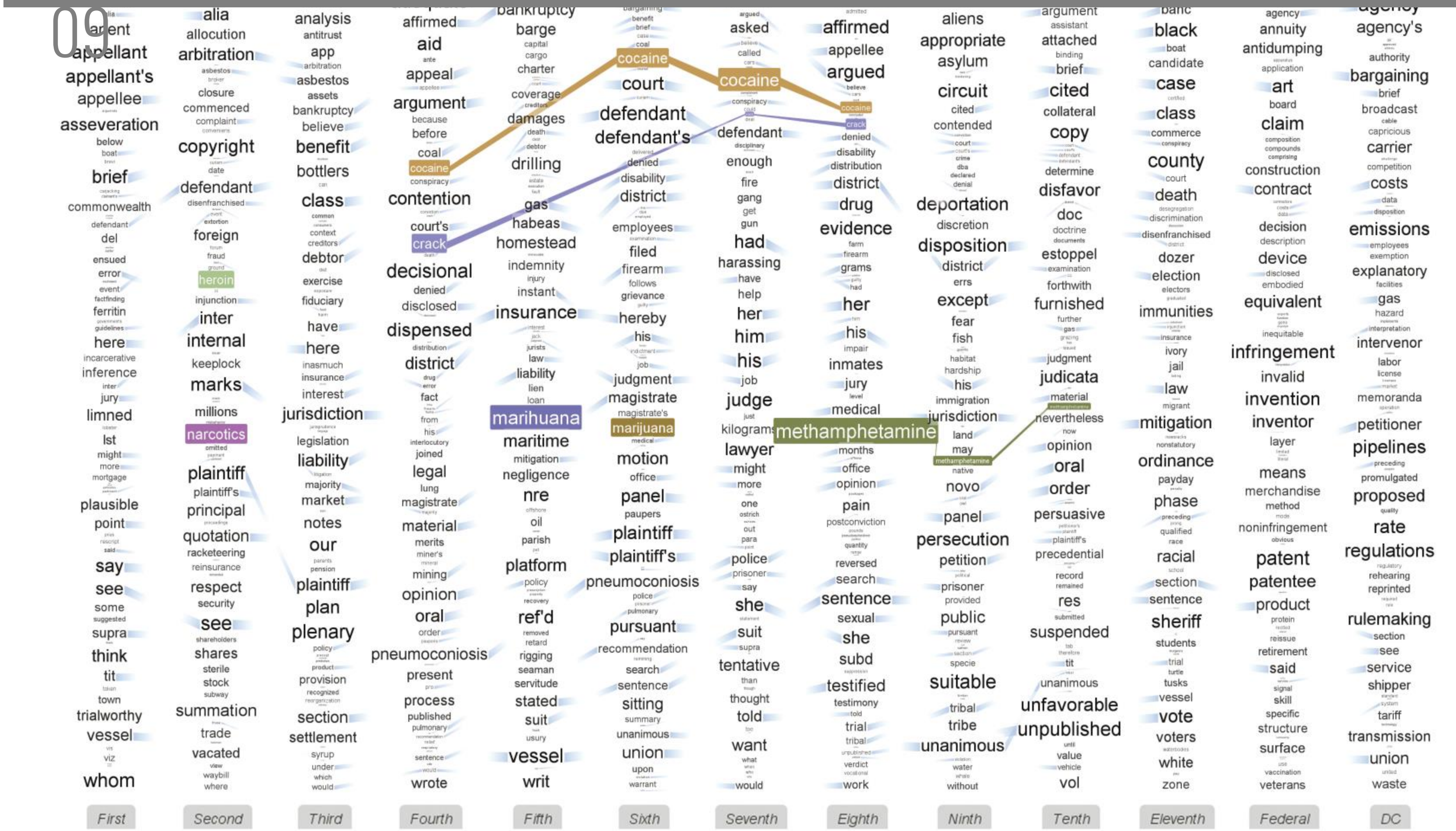

Improving the Readability of Clustered Social Networks using Node Duplication

Nathalie Henry, Anastasia Bezerianos, and Jean-Daniel Fekete

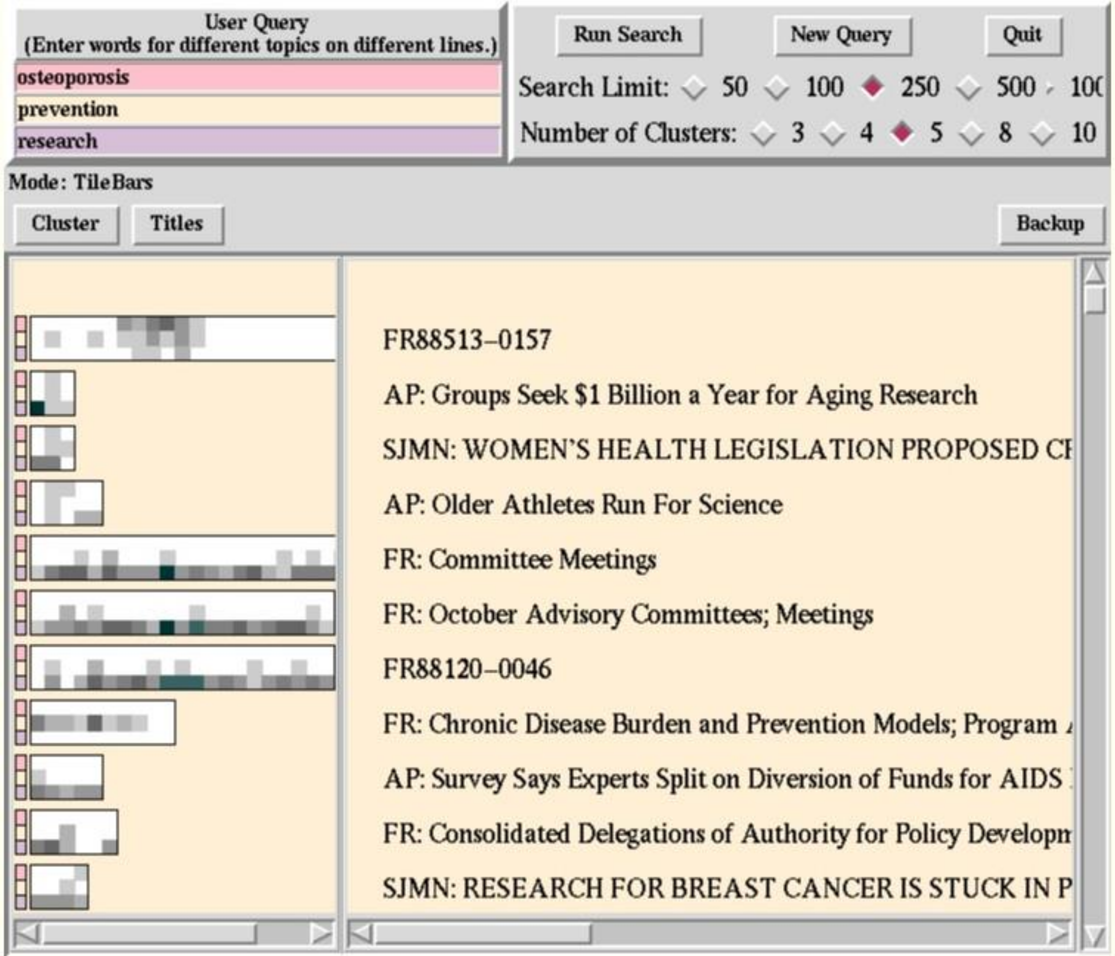
THEMERIVER HAVRE ET AL 1999



PARALLEL TAG CLOUDS



SUPPORTING SEARCH



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

osteoporosis
prevention
research

Mode: TileBars

FR88513–0157

AP: Groups Seek $1 Billion a Year for Aging Research

SJMN: WOMEN'S HEALTH LEGISLATION PROPOSED CF

AP: Older Athletes Run For Science

FR: Committee Meetings

FR: October Advisory Committees; Meetings

FR88120–0046

FR: Chronic Disease Burden and Prevention Models; Program

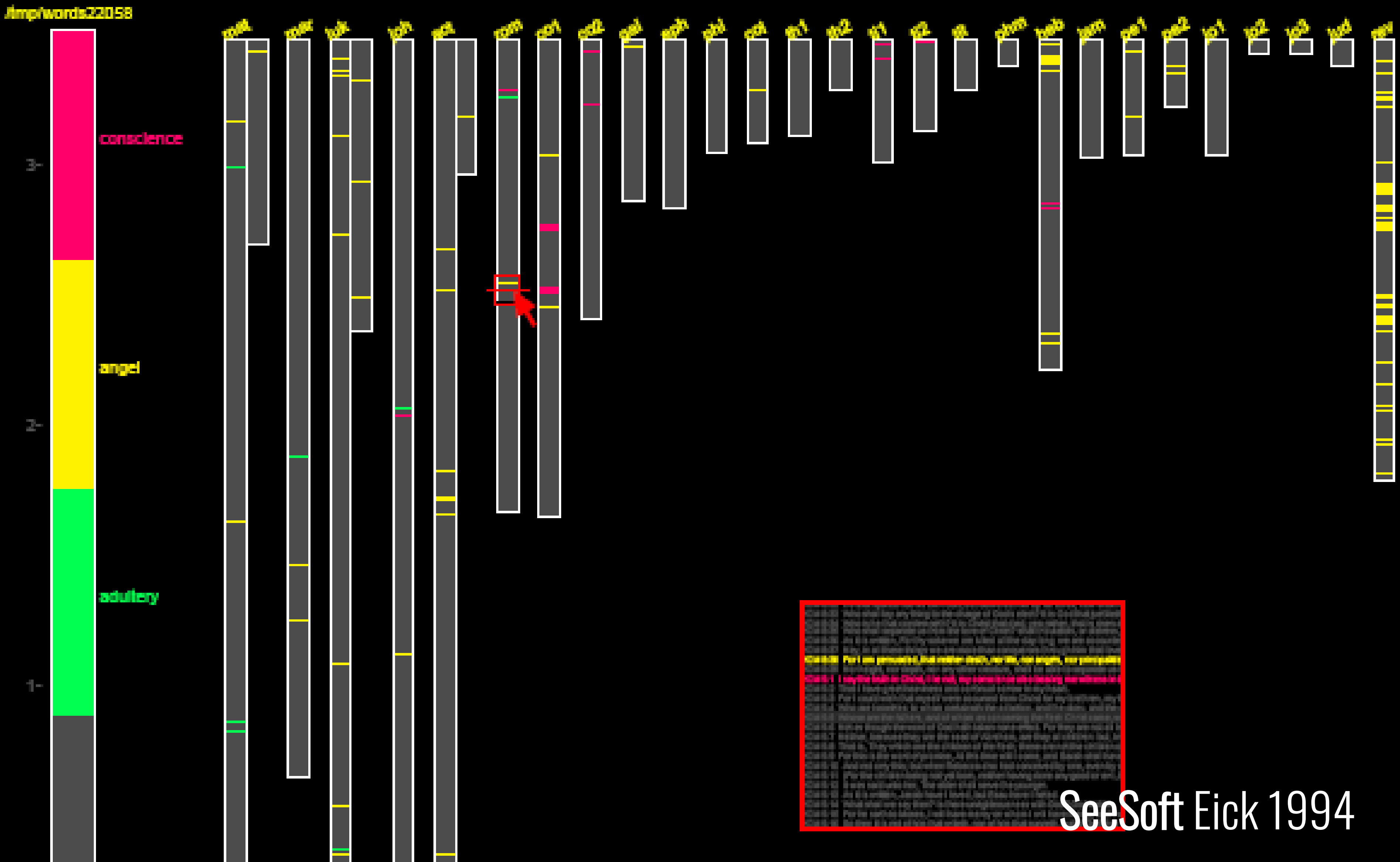
AP: Survey Says Experts Split on Diversion of Funds for AIDS

FR: Consolidated Delegations of Authority for Policy Developm

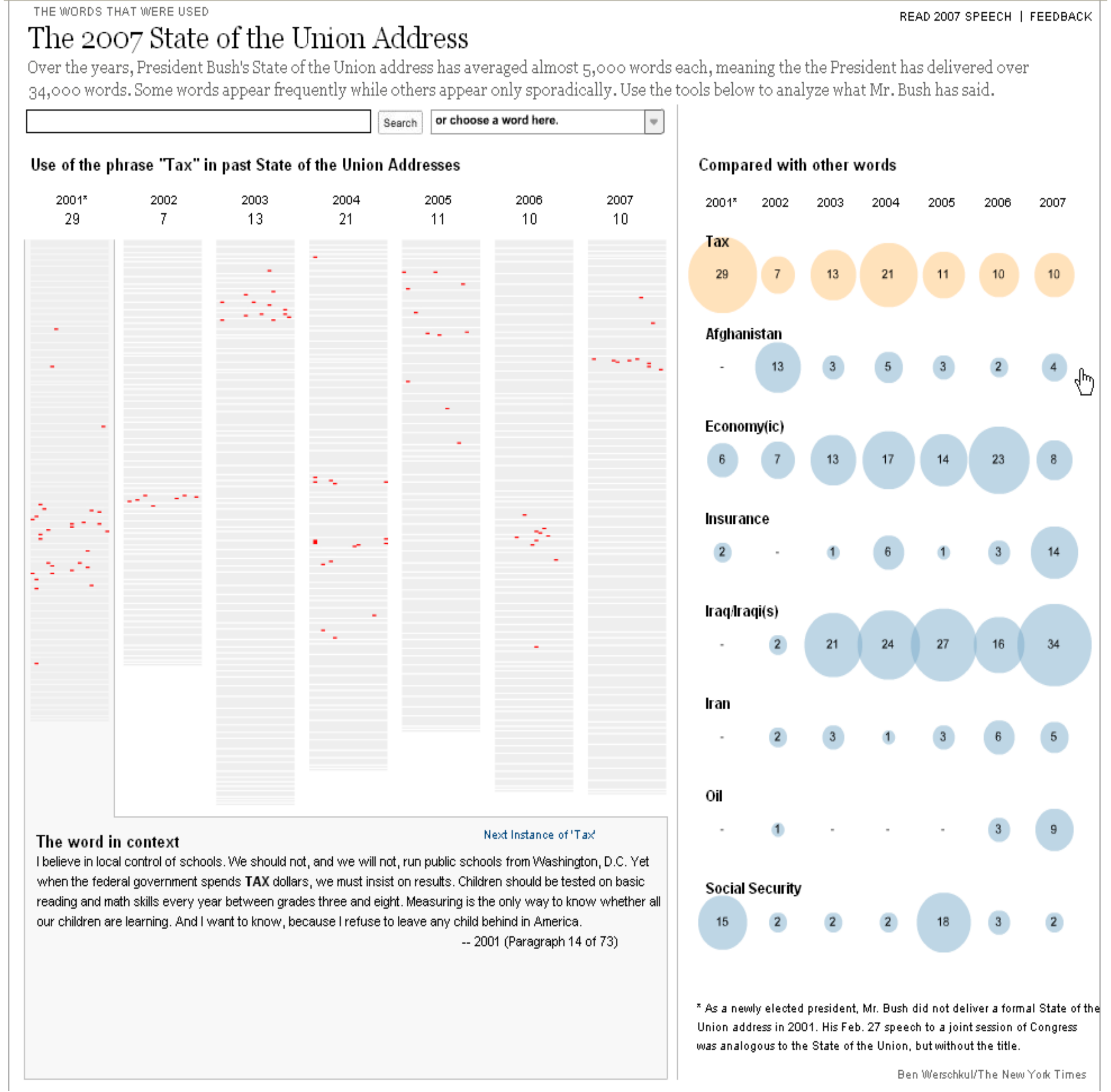
SJMN: RESEARCH FOR BREAST CANCER IS STUCK IN P

Run Search
New Query
Quit
Search Limit: ◇ 50 ◇ 100 ◆ 250 ◇ 500 ◇ 100
Number of Clusters: ◇ 3 ◇ 4 ◆ 5 ◇ 8 ◇ 10
Cluster
Titles
Backup

TileBars Hearst 1999



SeeSoft Eick 1994



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.

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

2001*	2002	2003	2004	2005	2006	2007
29	7	13	21	11	10	10

Compared with other words

	2001*	2002	2003	2004	2005	2006	2007
Tax	29	7	13	21	11	10	10
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

The word in context

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

-- 2001 (Paragraph 14 of 73)

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

Ben Werschkul/The New York Times

DOCUMENT SIMILARITY & CLUSTERING

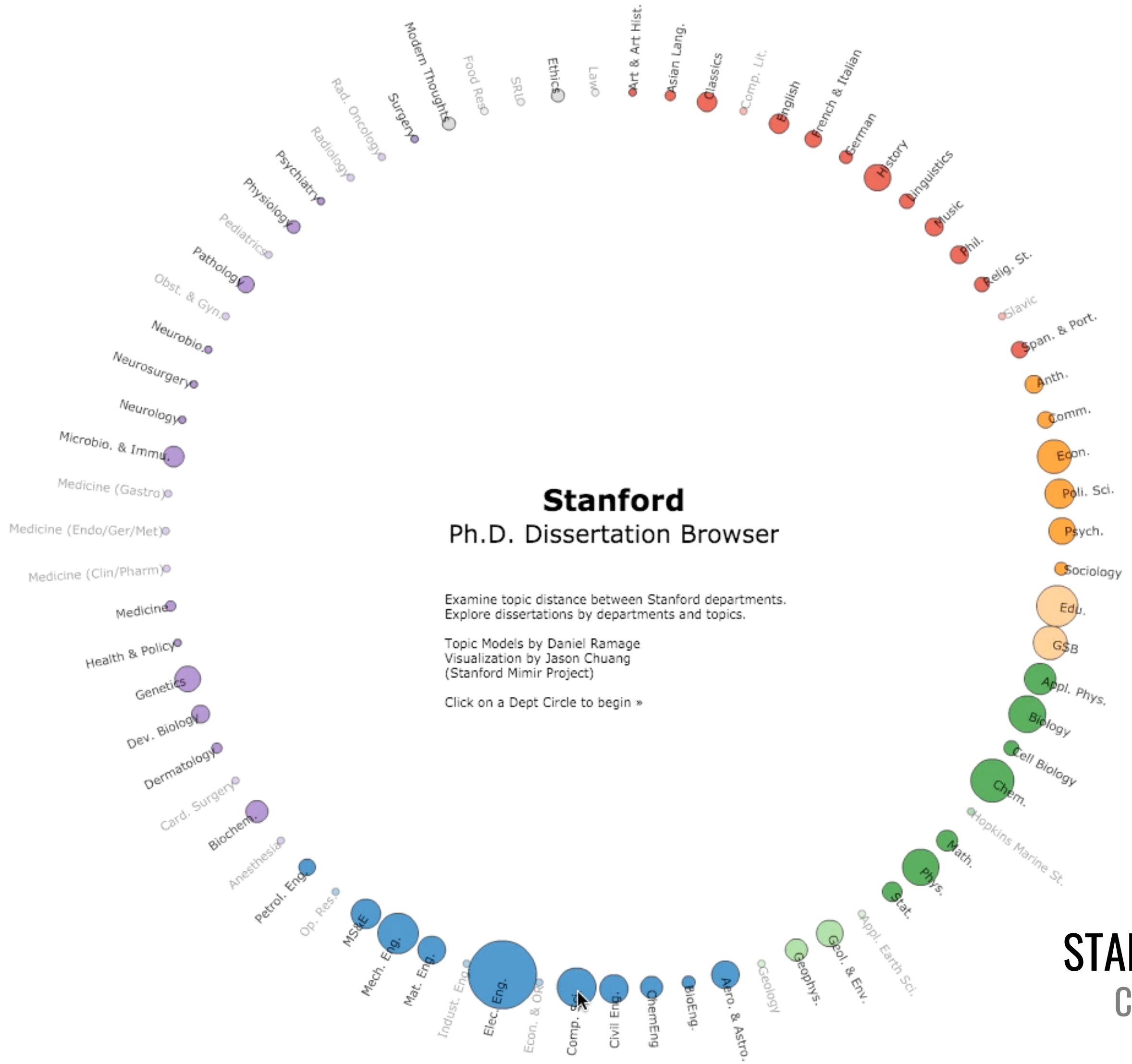
COMPUTE SIMILARITY BETWEEN DOCUMENTS BASED ON THE WORDS THEY SHARE

- **TF-IDF** (TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY) IS COMMON

TOPIC MODELING APPROACHES

- ASSUME DOCUMENTS ARE A MIXTURE OF TOPICS
- TOPICS ARE (ROUGHLY) A SET OF CO-OCCURRING TERMS
- LATENT SEMANTIC ANALYSIS (LSA): REDUCE TERM MATRIX

- MANY, MANY APPROACHES EXIST



Stanford

Ph.D. Dissertation Browser

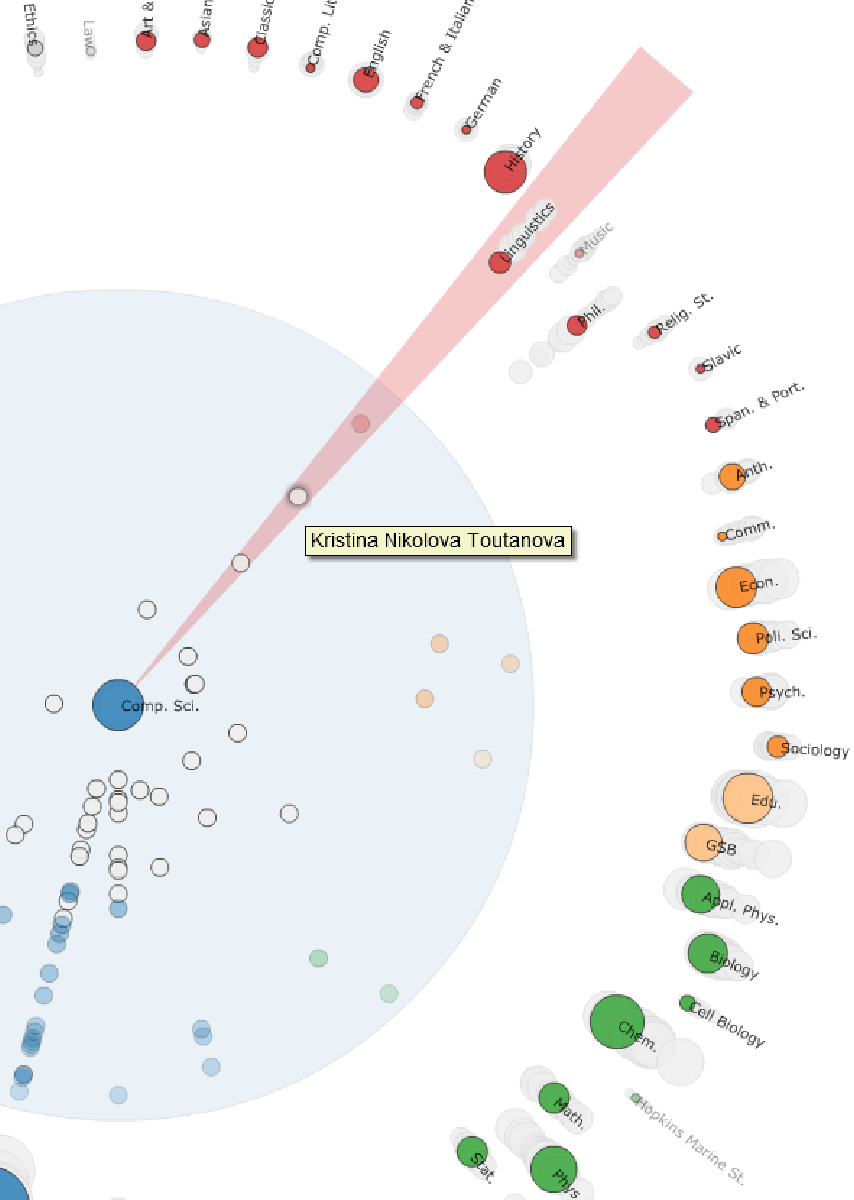
Examine topic distance between Stanford departments.
Explore dissertations by departments and topics.

Topic Models by Daniel Ramage
Visualization by Jason Chuang
(Stanford Mimir Project)

Click on a Dept Circle to begin »

STANFORD DISSERTATION BROWSER

CHUANG, RAMAGE, MANNING & HEER 2012



Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

Computer Science (2005)

Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

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

STANFORD DISSERTATION BROWSER

CHUANG, RAMAGE, MANNING & HEER 2012

WARNING

OFTEN, TEXT VISUALIZATIONS DO NOT REPRESENT TEXT DIRECTLY, BUT THEY REPRESENT A MODEL

WORD COUNTS, WORD SEQUENCES, CLUSTERS, ETC.

ASK:

CAN YOU INTERPRET THE VISUALIZATION?

DOES THE MODEL ACCURATELY REPRESENT THE ORIGINAL TEXT?

LESSONS FOR TEXT VISUALIZATION

SHOW SOURCE TEXT (OR PROVIDE ACCESS TO IT)

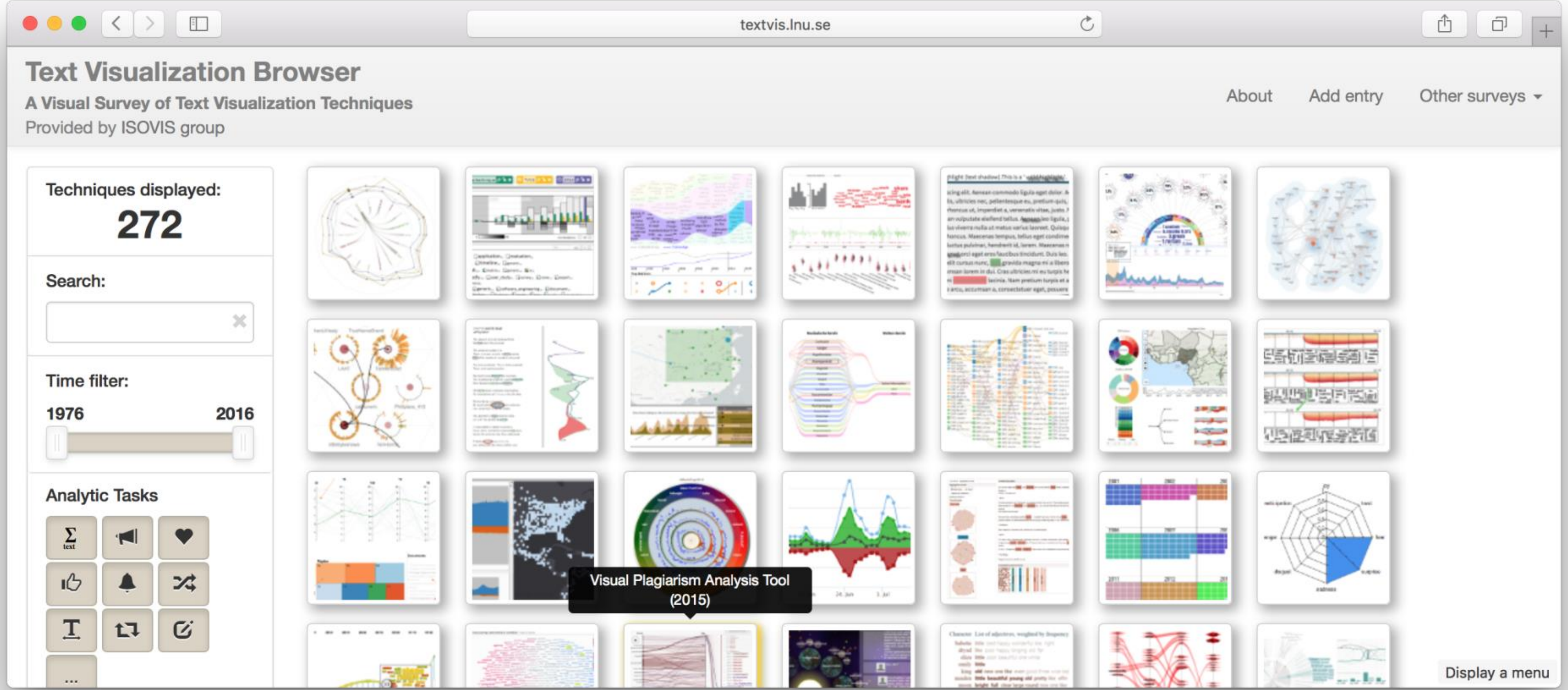
WHERE POSSIBLE, USE VISUALIZATION AS INDEX INTO DOCUMENTS

GROUP DOCUMENTS IN MEANINGFUL WAYS

WILL VIEWERS UNDERSTAND THE CLUSTERS?

WHERE POSSIBLE **USE TEXT TO REPRESENT TEXT**

HUNDREDS OF TOOLS & TECHNIQUES FOR TEXT AT http://textvis.lnu.se/



textvis.lnu.se

Text Visualization Browser

A Visual Survey of Text Visualization Techniques

Provided by ISOVIS group

About · Add entry · Other surveys ▾

Techniques displayed: 272

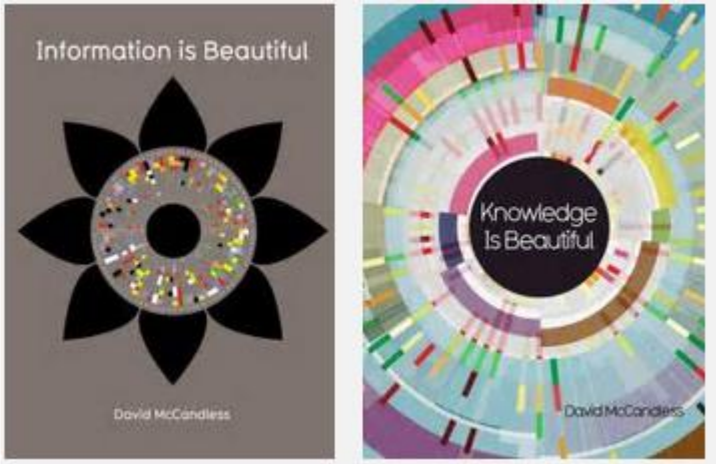
Search:

Time filter: 1976 – 2016

Analytic Tasks

HTTPS://YOUTU.BE/HOEX1IU57IW?T=420

Getting to know David McCandless



HTTP://WWW.WEFEELFINE.ORG/

QUESTIONS?

ACKNOWLEDGEMENTS

Slides in were inspired, adapted, taken from slides by

- Christopher Collins (University of Ontario Institute of Technology)
- Wesley Willett (University of Calgary)