VISUALIZING TEXT

Petra Isenberg

informatics mathematics
Inria

RECAP

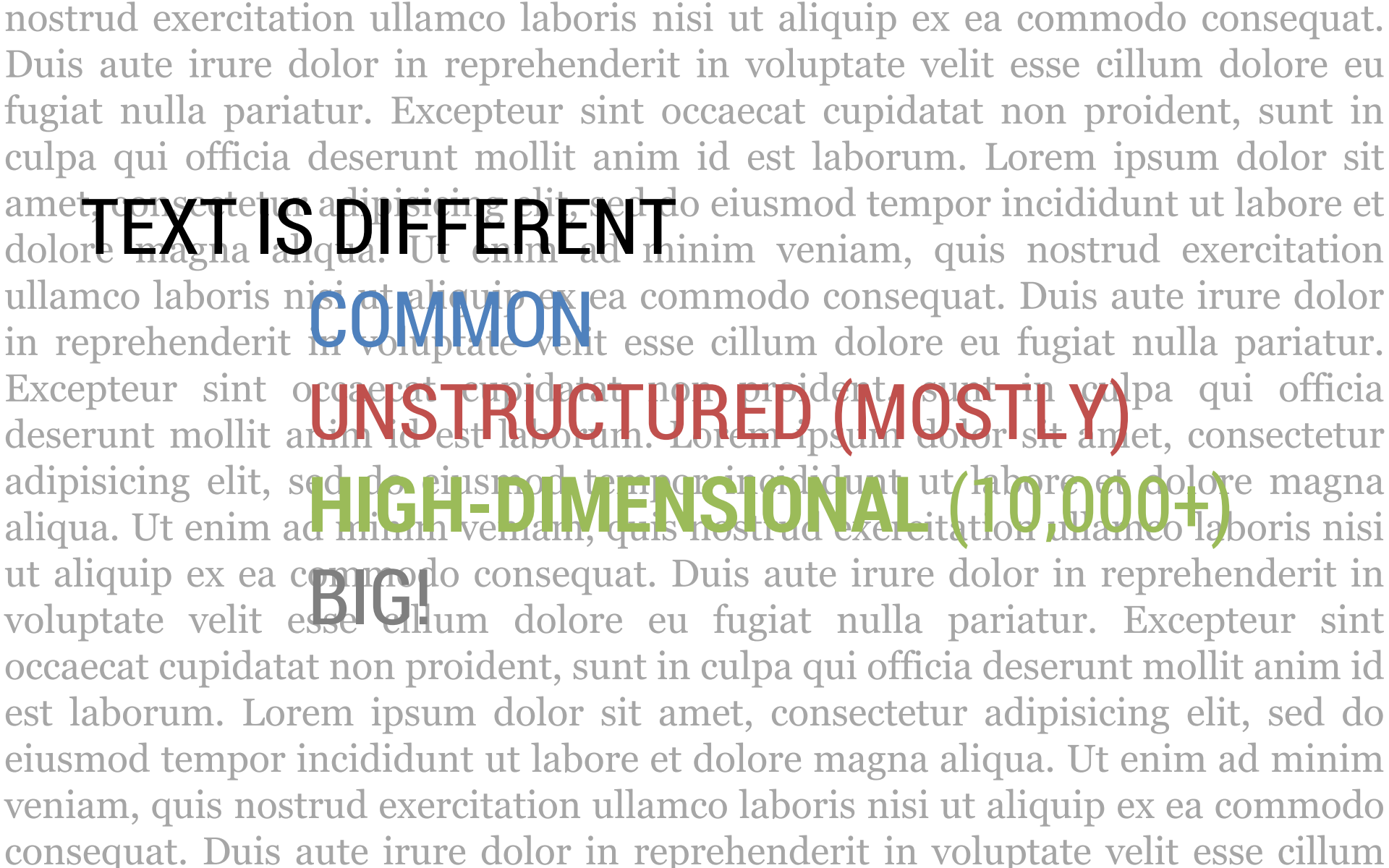
STRUCTURED DATA

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)

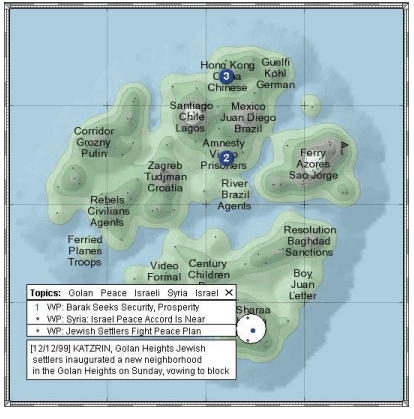
VISUALIZING TEXT


TEXT IS DIFFERENT
COMMON
UNSTRUCTURED (MOSTLY)
HIGH-DIMENSIONAL (10,000+)
BIG!

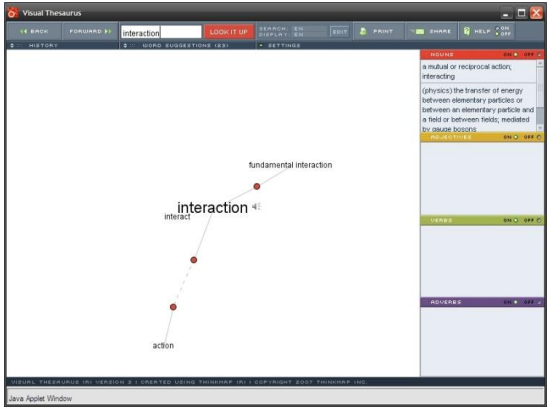
TEXT?

WHY

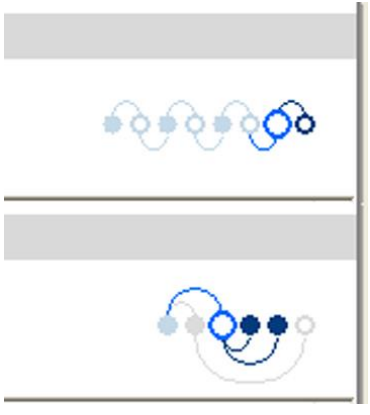
- To assist information retrieval
- To enable linguistic analysis
- To augment analytics on mixed data



Themescape



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:
  - "Tables encourage casual interaction."

  vs

  - "I encouraged her to take a day off."

LANGUAGE IS AMBIGUOUS

- I saw Pathfinder on Mars with a telescope.
- Pathfinder photographed Mars.
- The Pathfinder photograph mars our perception of a lifeless planet.
- The Pathfinder photograph from Ford has arrived.
- The Pathfinder forded the river without marring its paint job.

(Hearst, 2006)

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

-

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

REPETITION
PLAGARISM
SHARED
ENTITIES
AUTHOR
STYLE

COLLECTION

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

TENSE
SENTIMENT
SENTENCE
LENGTH
READING

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

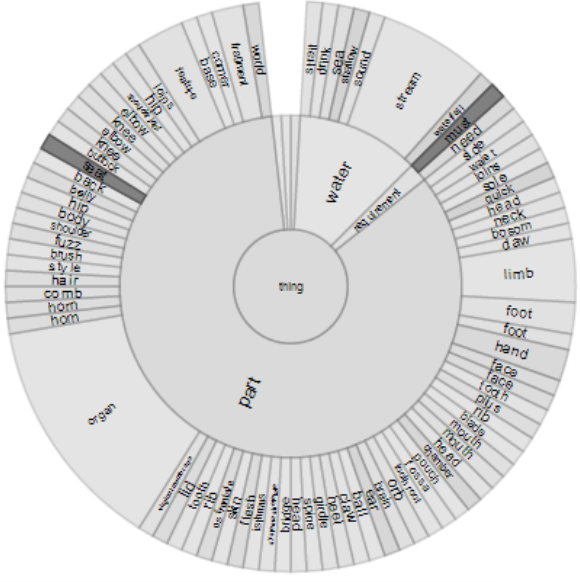
- A good way to increase signal and reduce fracturing of the corpus if there aren’t many words.
- Note: Keep the original words somewhere! Also keep the case if you choose to lowercase the word; you never know when you’ll need this data

STOP WORD REMOVAL

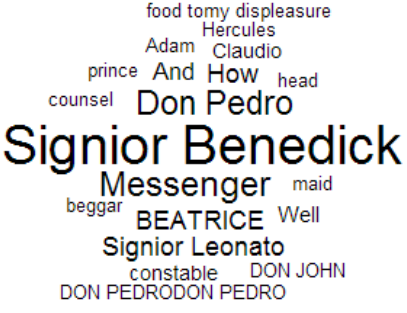
- Common words such as “and”, “the”, “I” are removed from view to highlight content words
- Domain specific stop words, e.g. in legal domain:
  - Court, attorney, honour, plaintiff, etc.
- Caution! These words have been shown to be useful for stylistic analysis! When working with text corpora, KEEP EVERYTHING.

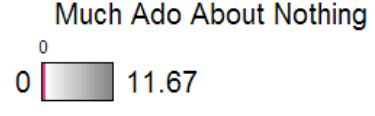
NAMED ENTITY RECOGNITION

- 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]

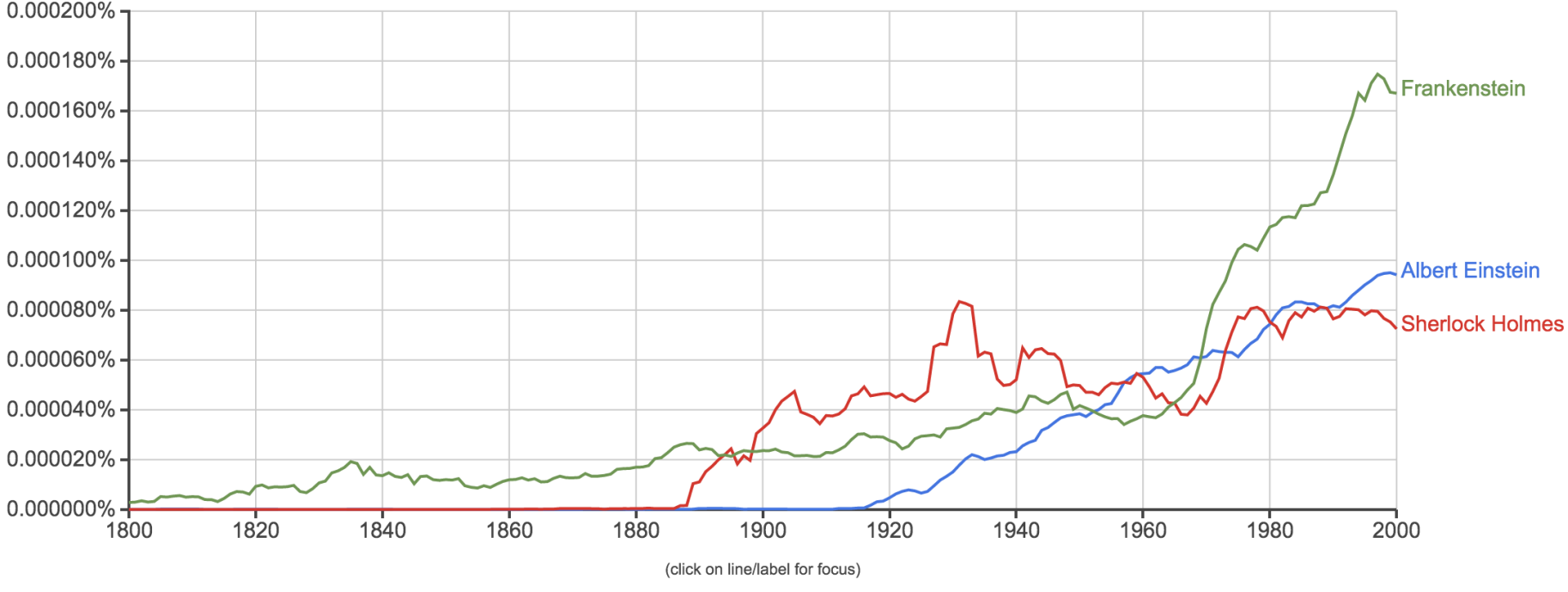
TOKENIZE (N=2)
[the quick], [quick brown], [brown fox], [fox jumps], [jumps over], [over the]…

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. Search lots of books



(click on line/label for focus)

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')]
```

NLTK.org
Python

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'), ('.', '.')])
```

DOCUMENT CONTENT

SKETCHING

SKETCHING: **VISUALIZE**

IMAGINE YOU HAVE 20 YEARS OF UNIVERSITY PH.D. THESES:

- YEAR
- DEPARTMENT
- AUTHOR
- ADVISOR
- COMMITTEE
- **COMPLETE TEXT**

TASK:

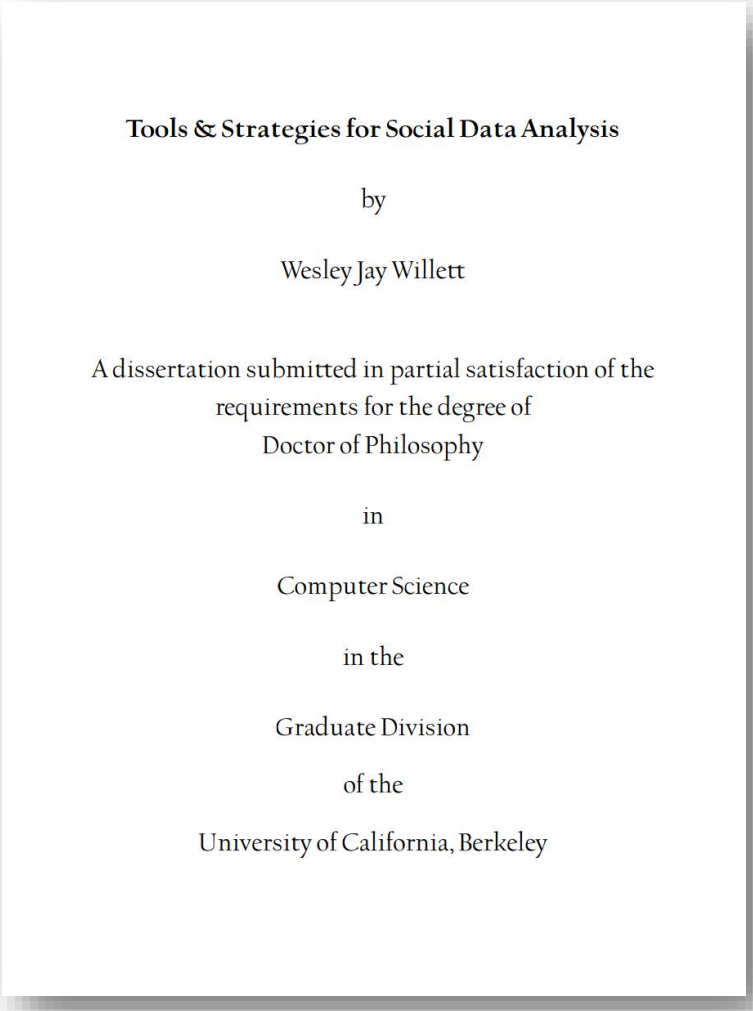
1) VISUALIZE THE MOST IMPORTANT CONTENT FROM A SINGLE THESIS.

2) VISUALIZE HOW SIMILAR THESES FROM EACH DEPARTMENT ARE TO THESES FROM OTHER DEPARTMENTS.

GROUPS OF 3

(~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?

additional air analysis analysts annotation applications approach asked author average based build chart citizen clustering collaborative collection comments commentspace community complete condition contributions crowd crowdsourcing data datasets design different discussion evidence example experiment experts explanations explore features figure filtering generated group help hypotheses hypothesis identify including indicating information interactive interface knowledge links members microtasks multiple novice number oaen observations organize participants phases pp proceedings process produced prompt provide quality questions rate redundant requires responses results score sense share showing similar site social source specific state strategies study support systems tags tasks tools understanding used users views visualization web work workers

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)

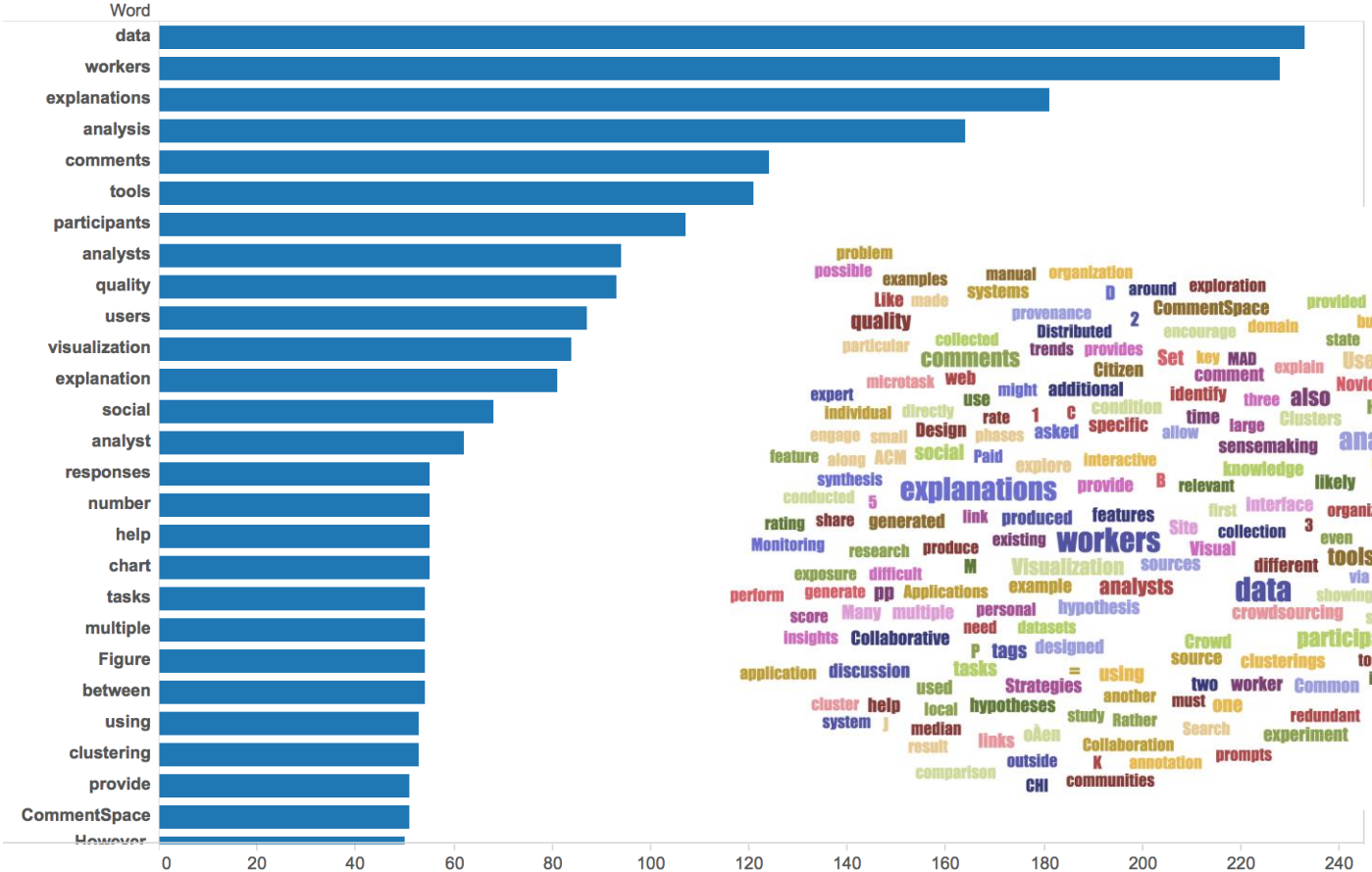
MAY NOT FACILITATE COMPARISON (UNSTABLE LAYOUT)

- 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
0
20
40
60
80
100
120
140
160
180
200
220
240


explanations
workers
data
analysis
tools
participants
analysts
visualization
comments
quality
also
explanation
Visualizations
Users
Figure
clustering
Community
Proceedings
CommentSpace
Distributed
Citizen
Crowd
Strategies
Collaboration
Collaborative
Applications
Monitoring
Approach
Information
Understanding
Sensing
Common
Search
Rather
Many
Like
Set
Novice
However
Responses
Experts
Clusters
Sensemaking
Knowledge
Interface
Organize
Findings
Discussions
Crowdsourcing
Clusterings
Commenting
Interviews
Observed
Process
Based
Evidence
Number
Support
Several
Prompts
Annotation
Communities
Hypotheses
Hypothesis
Explore
Interactive
Provide
Produced
Features
Generated
Existing
Visual
Sources
Example
Personal
Datasets
Designed
Tags
Tasks
Using
Another
Redundant
Experiment
Outside
Links
Median
Local
Cluster
System
Insights
Exposure
Difficult
Generate
Perform
Rating
Share
Link
Paid
Phases
Asked
Specific
Allow
Large
Time
Identify
Three
Comment
Explain
Key
MAD
Trends
Provides
Provenance
Encourage
Domain
Build
Good
View
Gathering
Whether
Provided
Around
Exploration
Organization
Manual
Systems
Examples
Possible
Problem
Across
Framework
Allows
Make
Analyst
Views
New
Task
Science
Within
Including
Focus
Observations
Response
Complete
Complex
Questions
Conditions
Work
Group
Filtering
Content
Microtasks
Prompt
Dataset
Chapter
Page
Average
Figure
Single
Likely
Relevant
Tag
Similar
Sensemaking
Knowledge
Interface
Organize
Findings
Discussions
Responses
Without
Scores
Text
Members
Showing
Sense
Community
Shows
Charts
Environmental
Understanding
Observed
Interviews
Commenting
Together
Included
Sensing
Chart
Process
Based
Similarity
Evidence
Number
Support
Several
Prompts
Annotation
Communities
Hypotheses
Hypothesis
Explore
Interactive
Provide
Produced
Features
Generated
Existing
Visual
Sources
Example
Personal
Datasets
Designed
Tags
Tasks
Using
Another
Redundant
Experiment
Outside
Links
Median
Local
Cluster
System
Insights
Exposure
Difficult
Generate
Perform
Rating
Share
Link
Paid
Phases
Asked
Specific
Allow
Large
Time
Identify
Three
Comment
Explain
Key
MAD
Trends
Provides
Provenance
Encourage
Domain
Build
Good
View
Gathering
Whether
Provided
Around
Exploration
Organization
Manual
Systems
Examples
Possible
Problem
Across
Framework
Allows
Make
Analyst
Views
New
Task
Science
Within
Including
Focus
Observations
Response
Complete
Complex
Questions
Conditions
Work
Group
Filtering
Content
Microtasks
Prompt
Dataset
Chapter
Page
Average
Figure
Single
Likely
Relevant
Tag
Similar
Sensemaking
Knowledge
Interface
Organize
Findings
Discussions
Responses
Without
Scores
Text
Members
Showing
Sense
Community
Shows
Charts
Environmental
Understanding
Observed
Interviews
Commenting
Together
Included
Sensing
Chart
Process
Based
Similarity
Evidence
Number
Support
Several
Prompts
Annotation
Communities

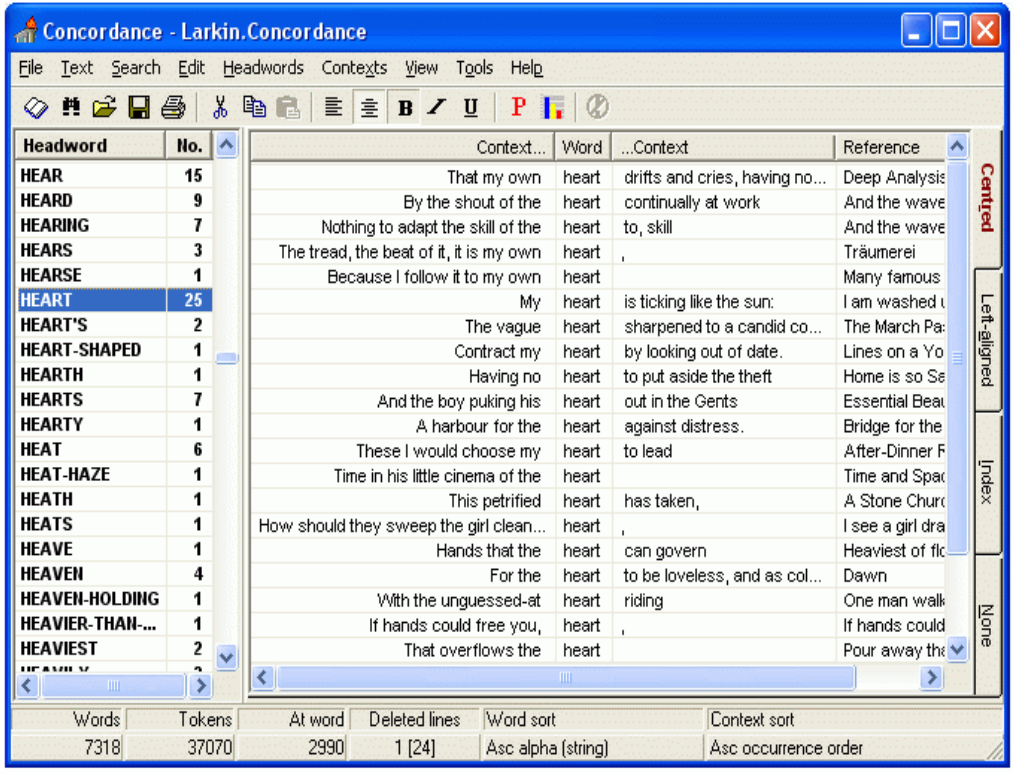
WORDCOUNT

http://wordcount.org

JONATHAN HARRIS

CONCORDANCE

WHAT IS THE COMMON LOCAL CONTEXT OF A TERM?



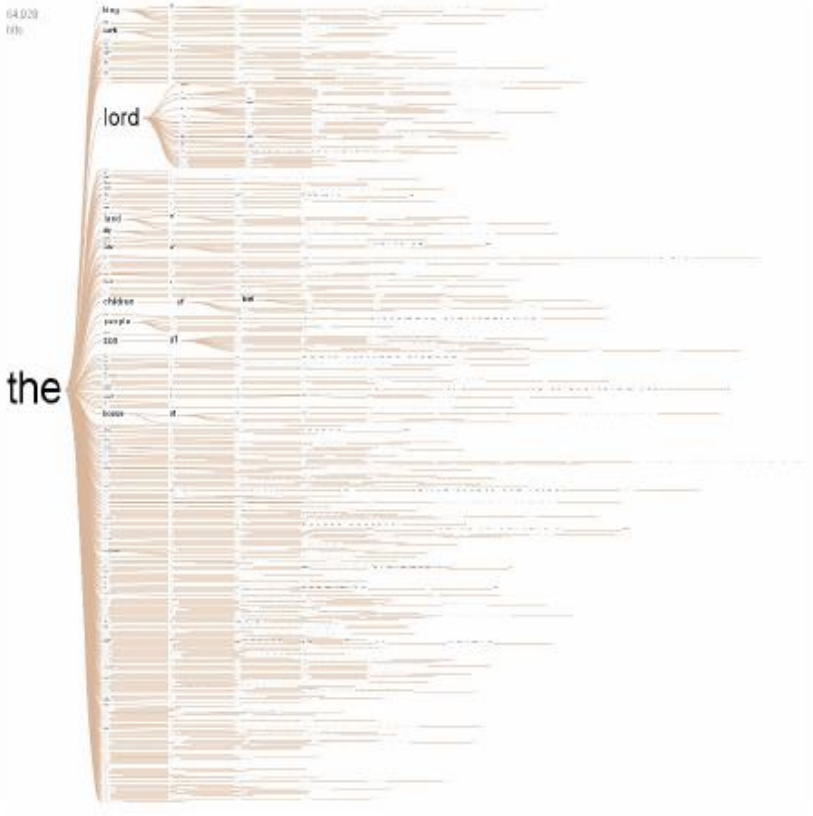
WORD TREES

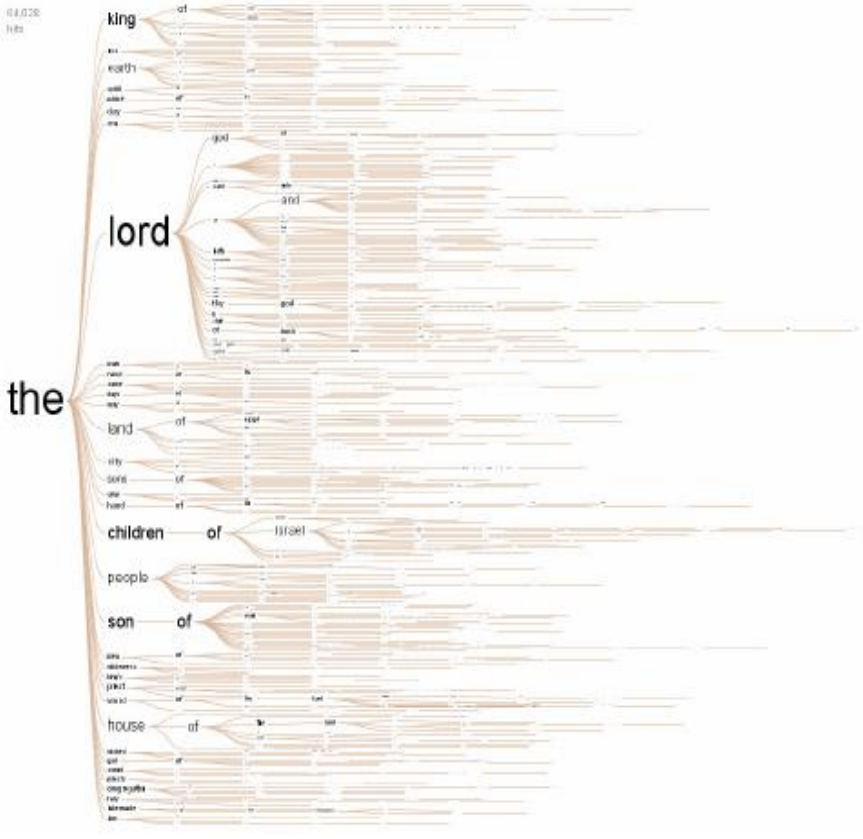
- cats are better than dogs
- cats eat kibble
- cats are better than hamsters
- cats are awesome
- cats are people too
- cats eat mice
- cats meowing
- cats in the cradle
- cats eat mice
- cats in the cradle lyrics
- cats eat kibble
- cats for adoption
- cats are family
- cats eat mice
- cats are better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

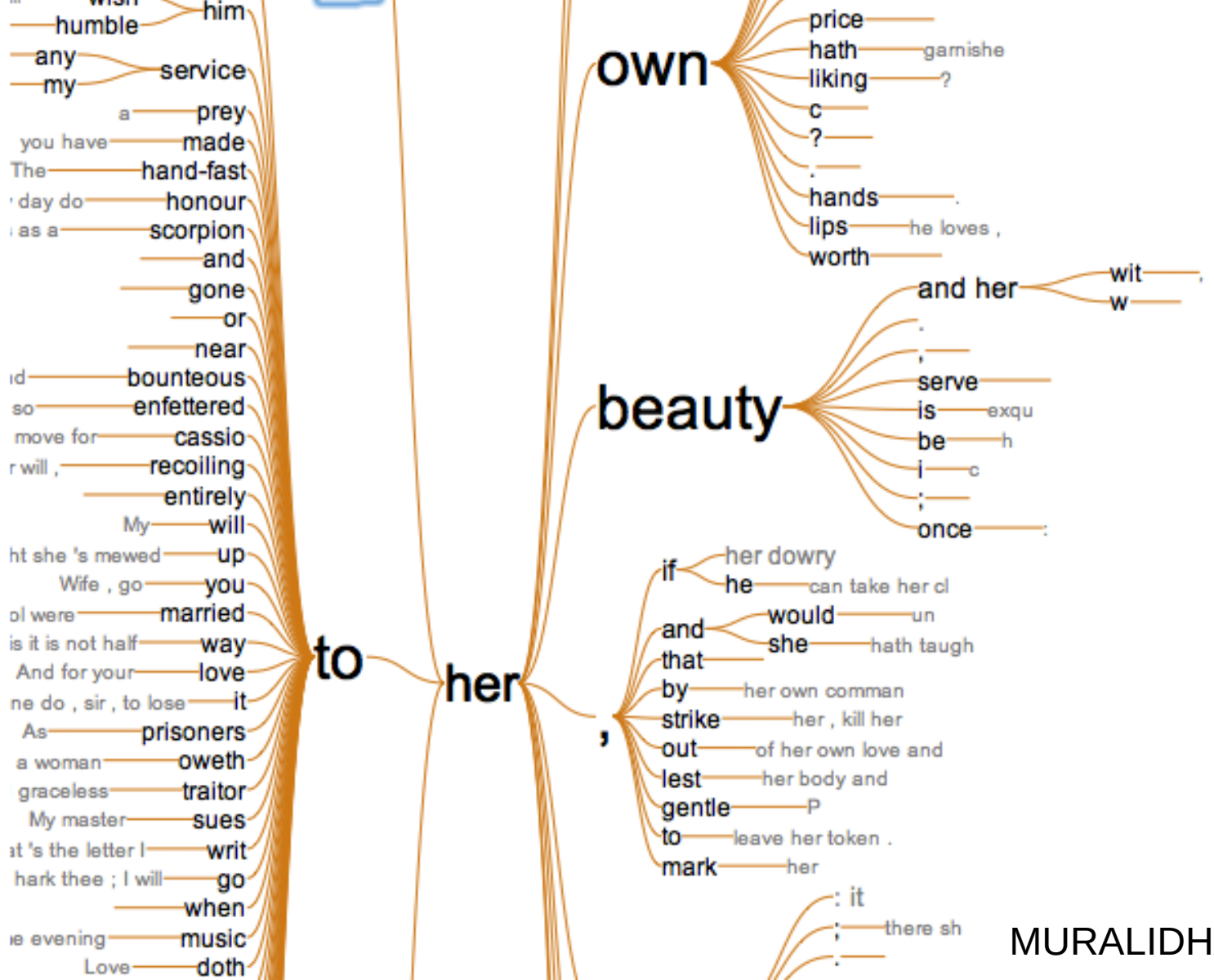
WATTENBERG & VIÉGAS 2008


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 lan
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



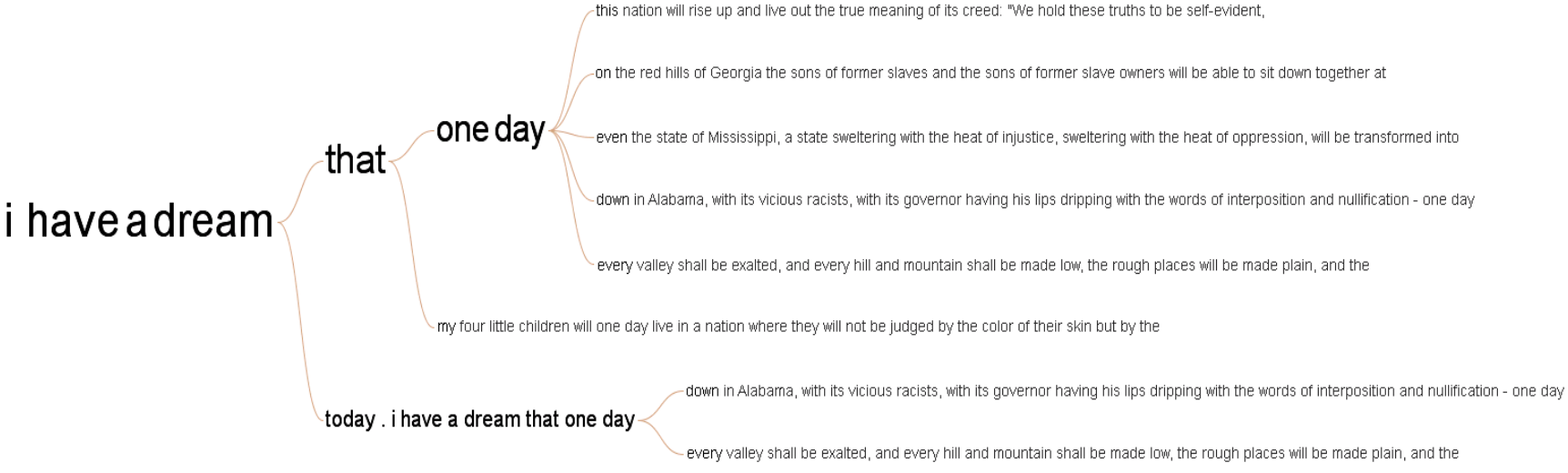




WORDSEER

MURALIDHARAN & HEARST

RECURRENT THEMES IN SPEECH



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

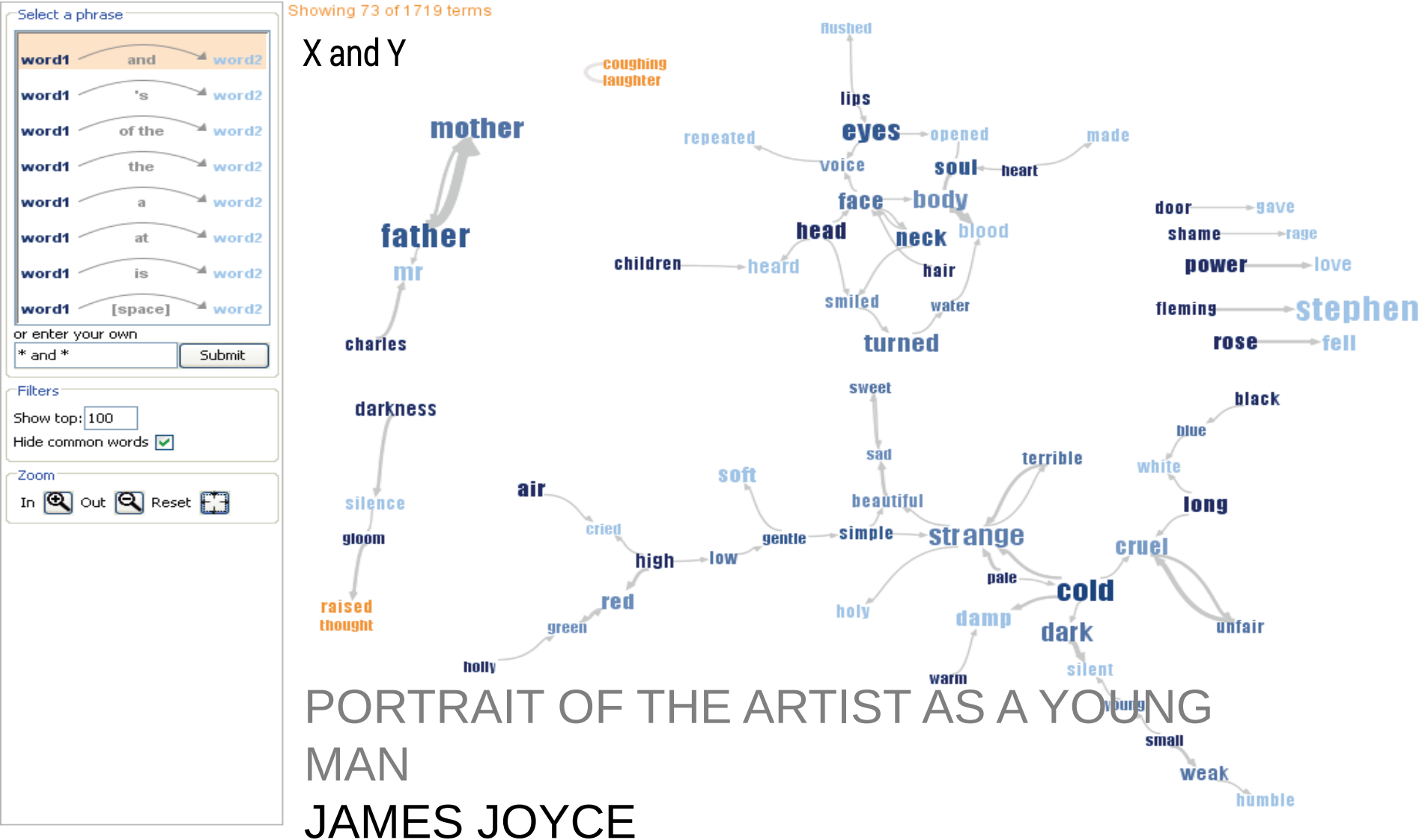
COULD BE OUTPUT OF REGEXP OR PARSER

VISUALIZE EXTRACTED PATTERNS IN A NODE-LINK VIEW

OCCURRENCES = NODE SIZE

PATTERN POSITION = EDGE DIRECTION

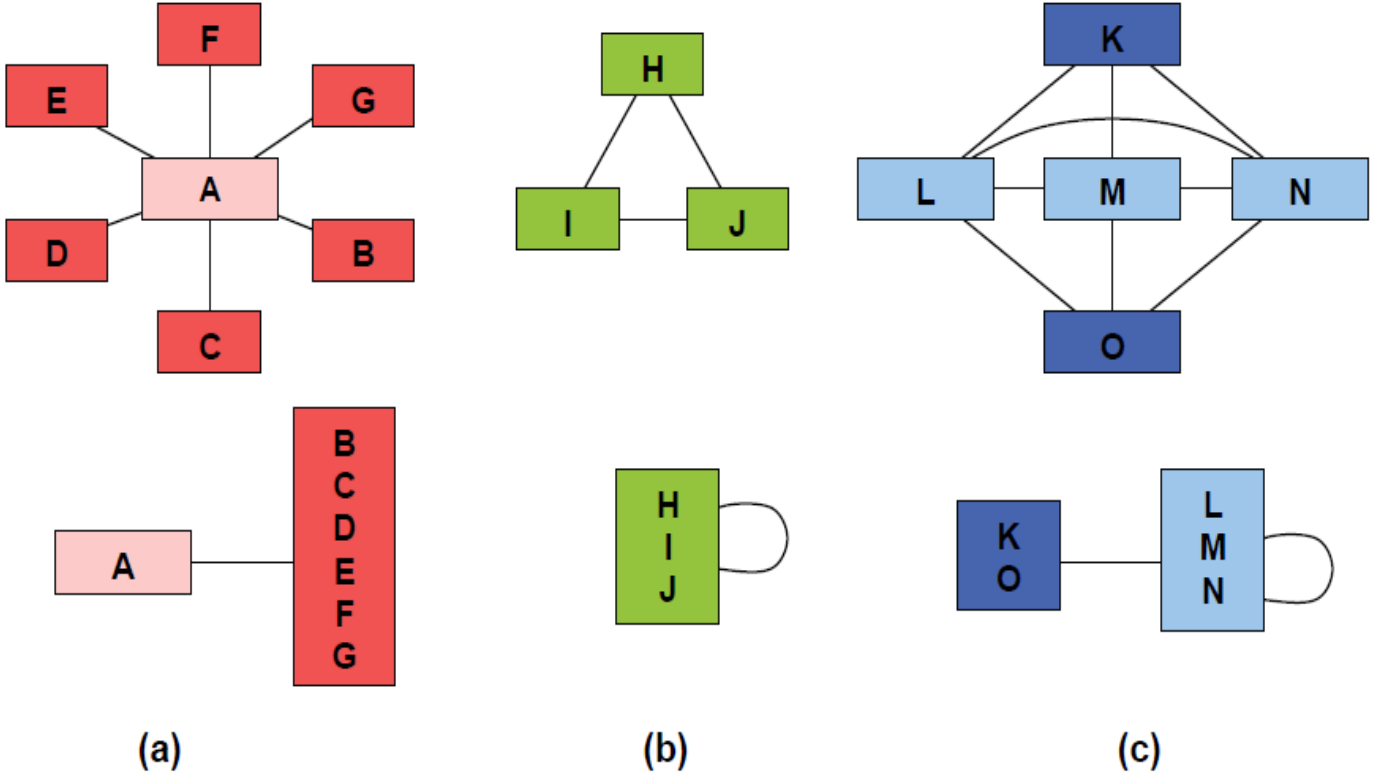
van Ham et al



PORTRAIT OF THE ARTIST AS A YOUNG MAN

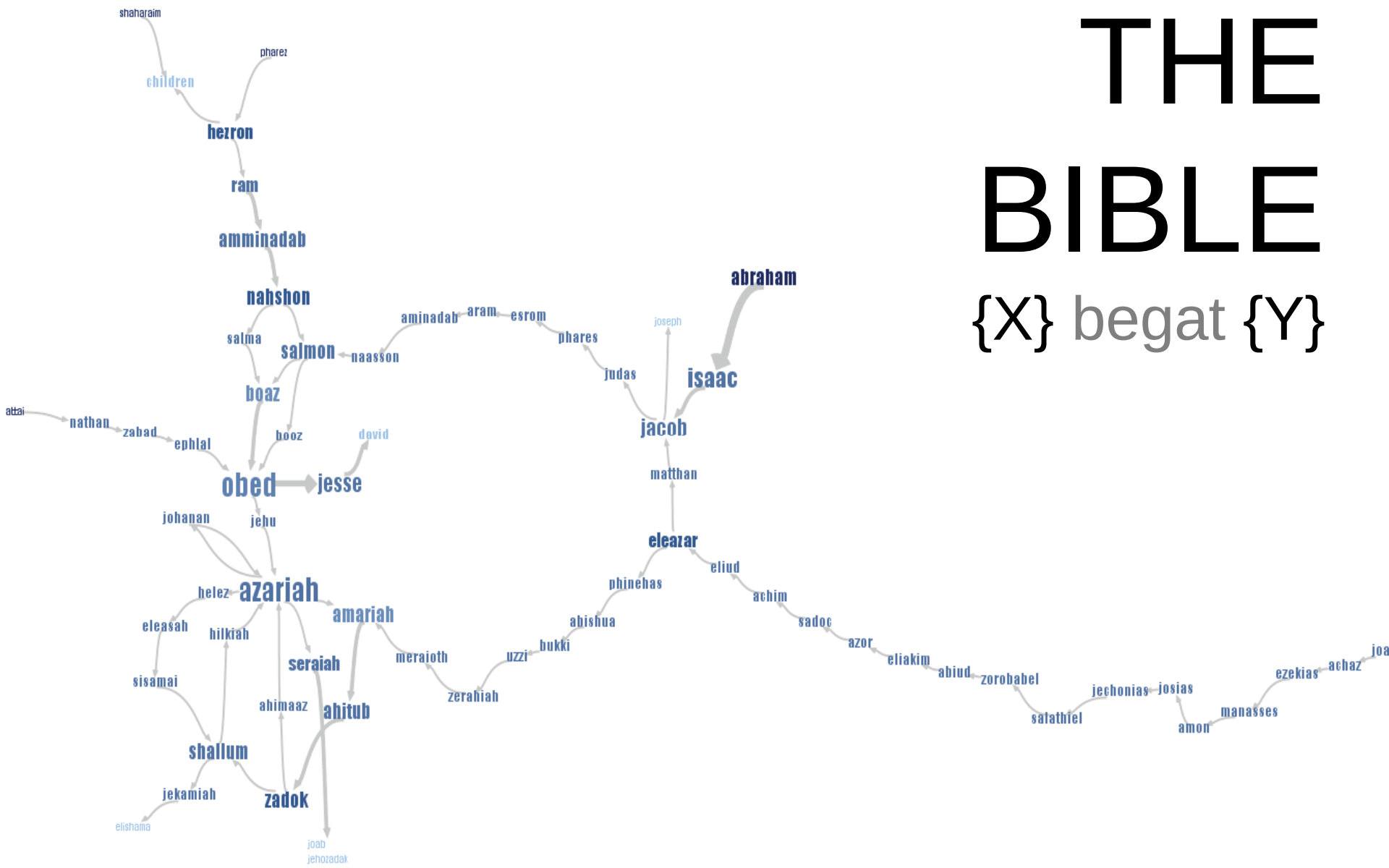
JAMES JOYCE

NODE GROUPING



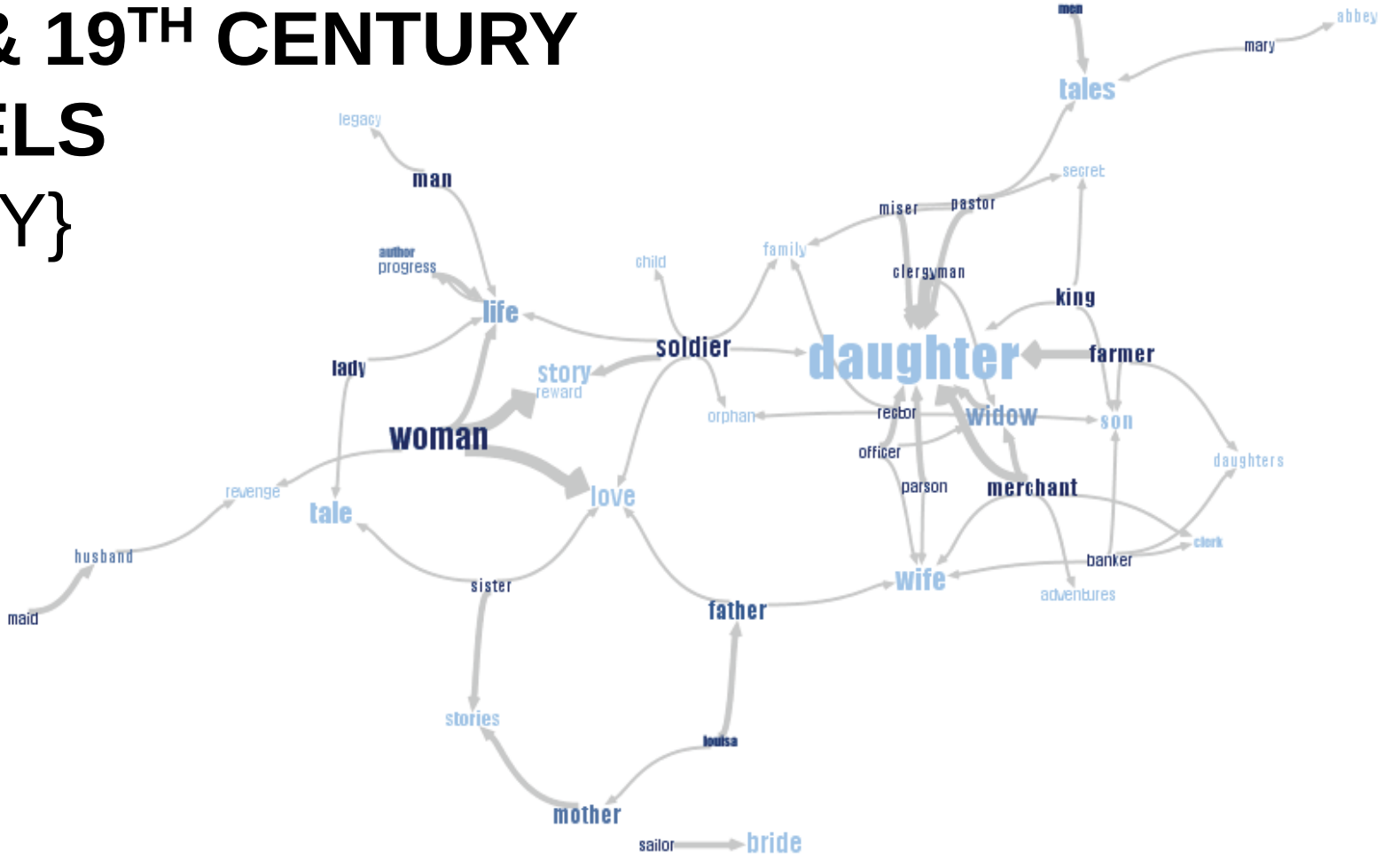
THE BIBLE

{X} begat {Y}



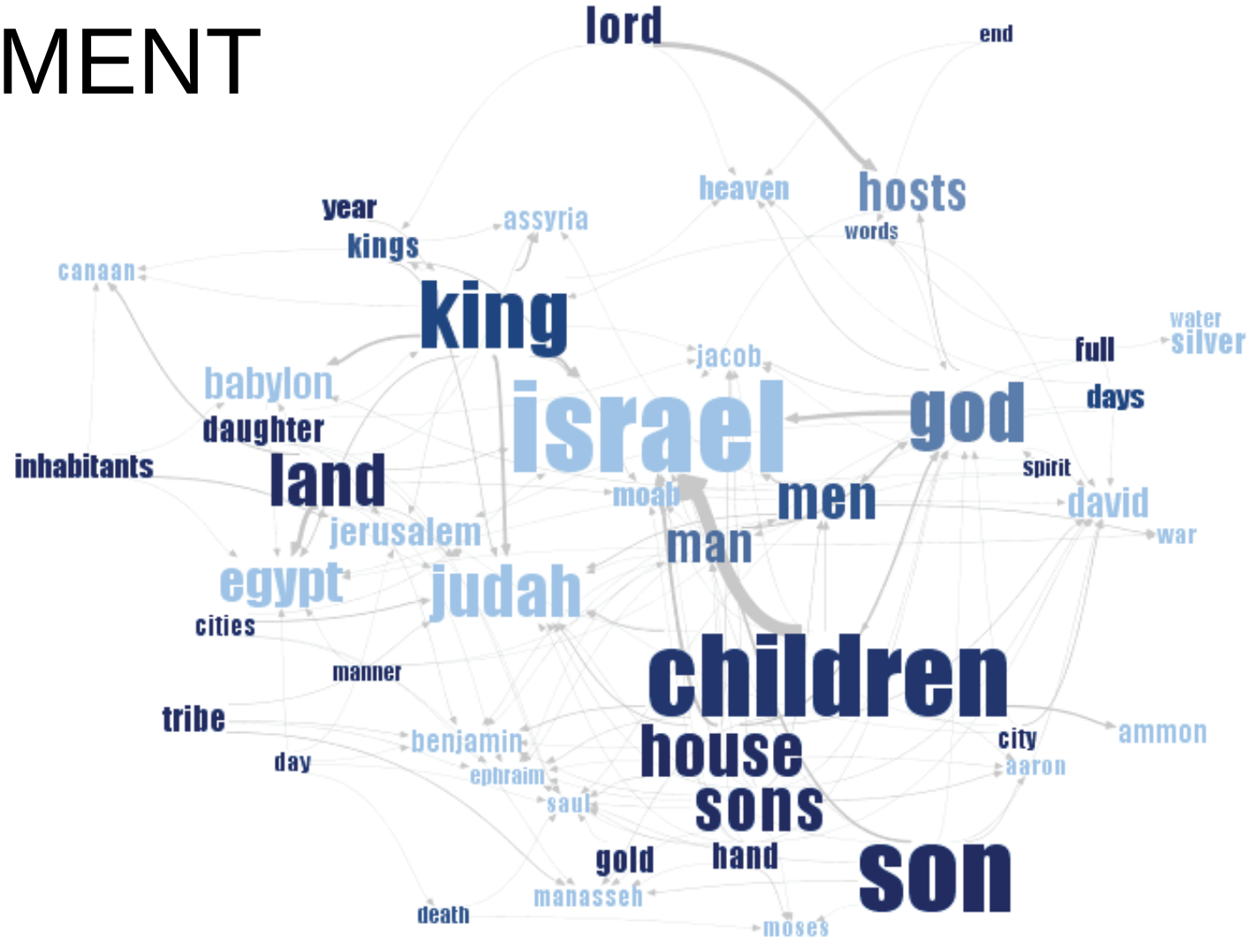
18TH & 19TH CENTURY NOVELS

{X}'s {Y}



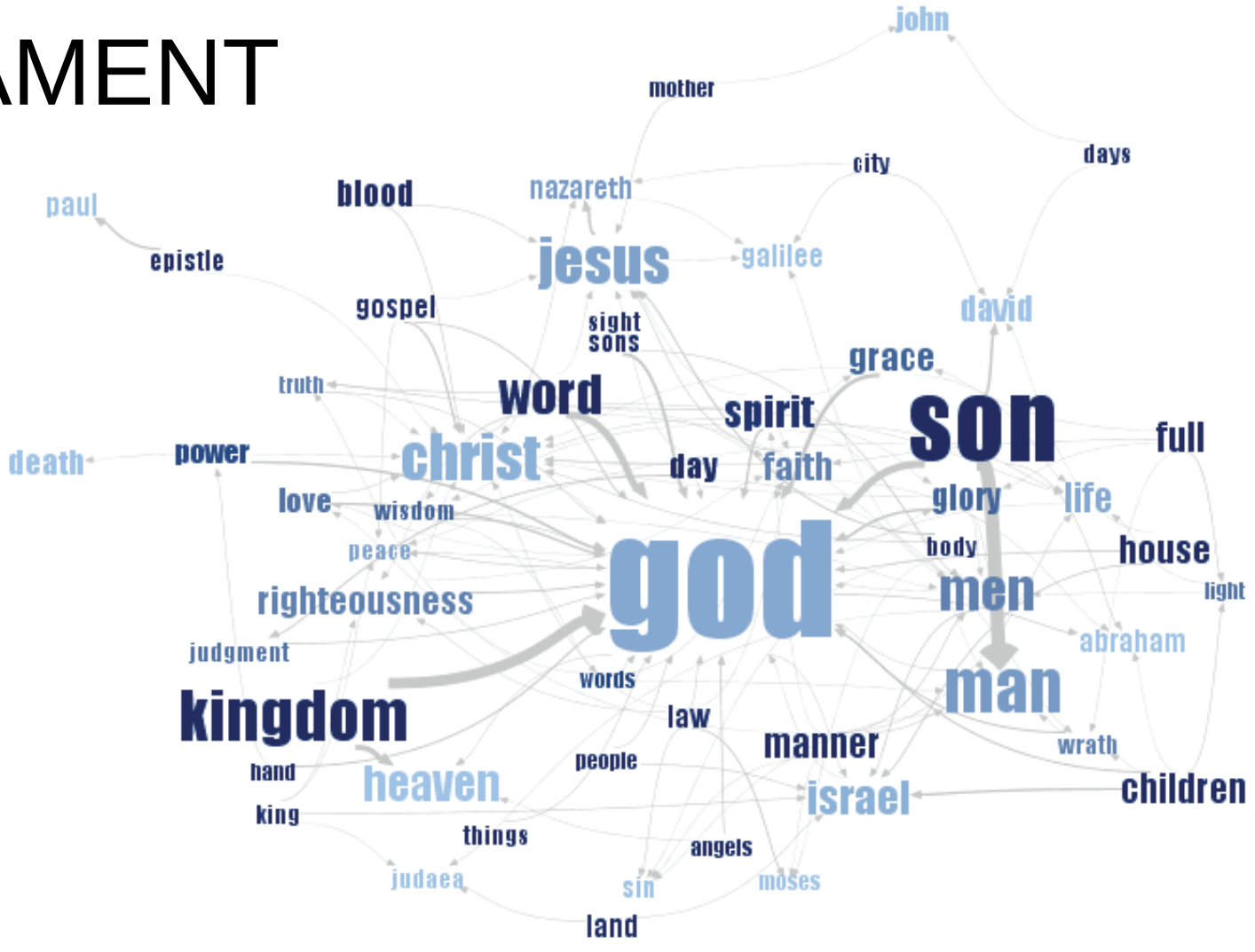
OLD TESTAMENT

{X} of {Y}



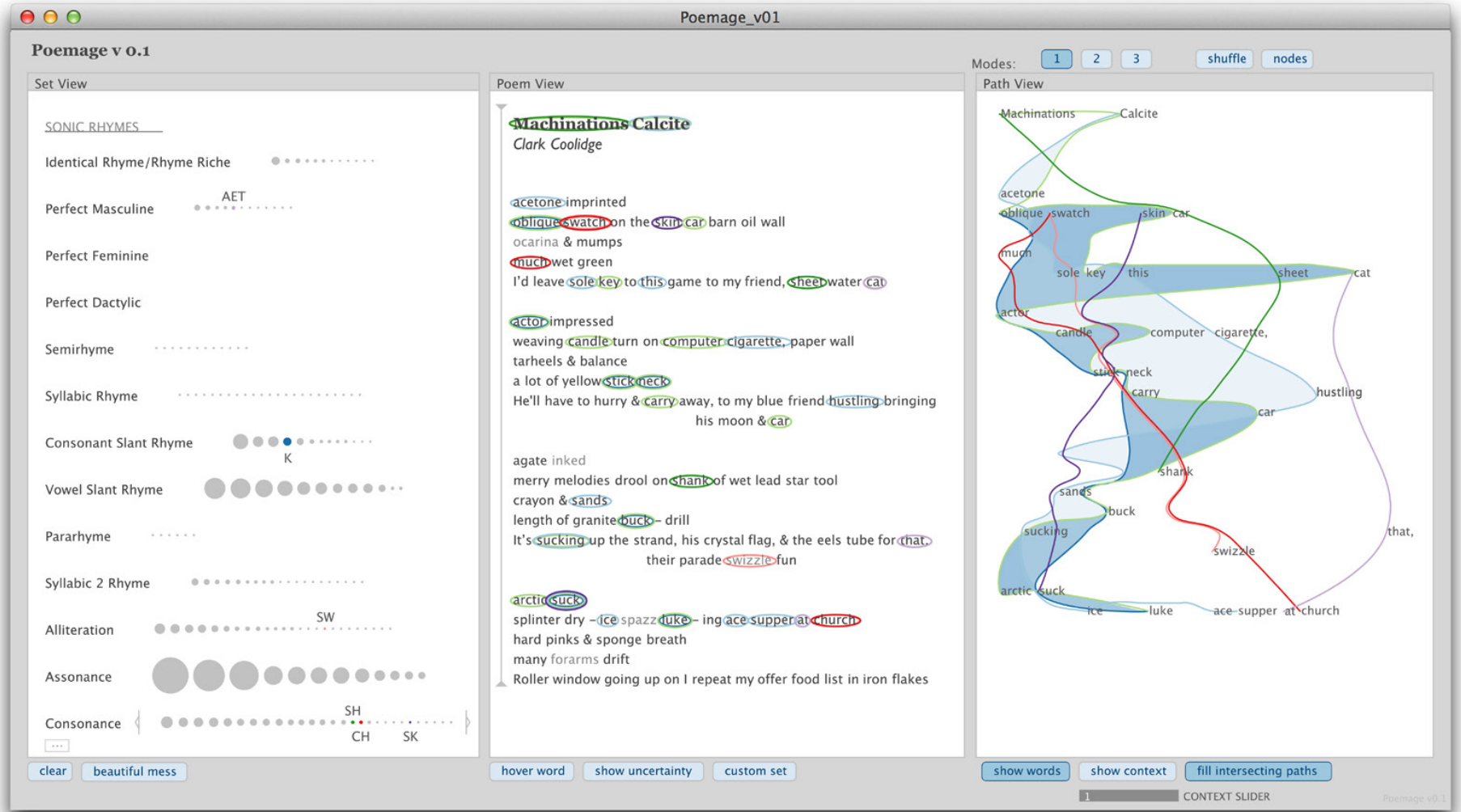
NEW TESTAMENT

{X} of {Y}



RHYME, SPEECH, ETC.

POEMAGE McCurdy et al. 2016



REVISIT YOUR SKETCHES?

TASK:

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

ARE YOUR VISUALIZATION CHOICES **EFFECTIVE**?

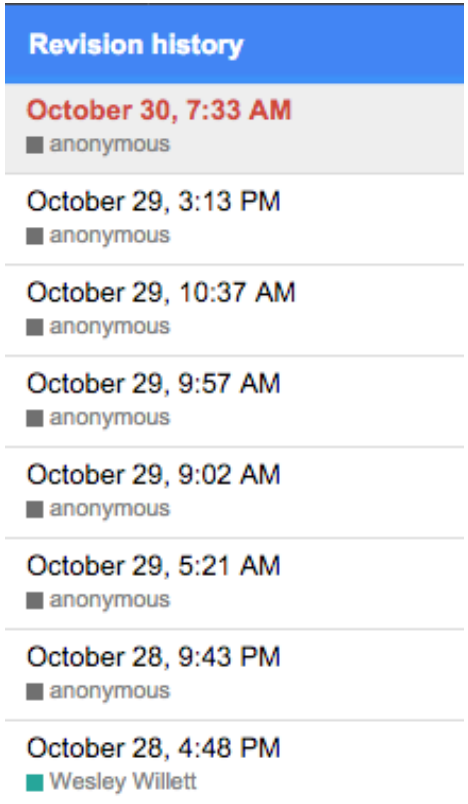
DOES THE VIS CAPTURE THE **LENGTH**, **FORM**, AND **POSITION** OF THE IMPORTANT CONTENT?

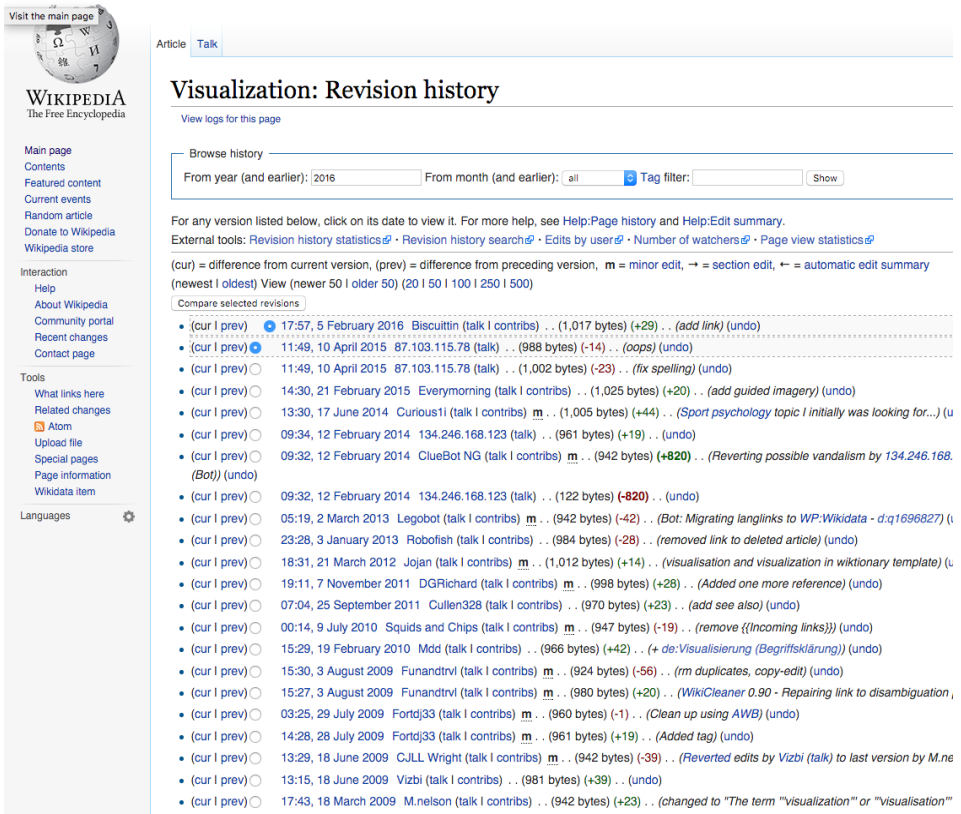
DO YOU SHOW OR CONNECT BACK TO THE **ORIGINAL TEXT**?

EVOLVING DOCUMENTS

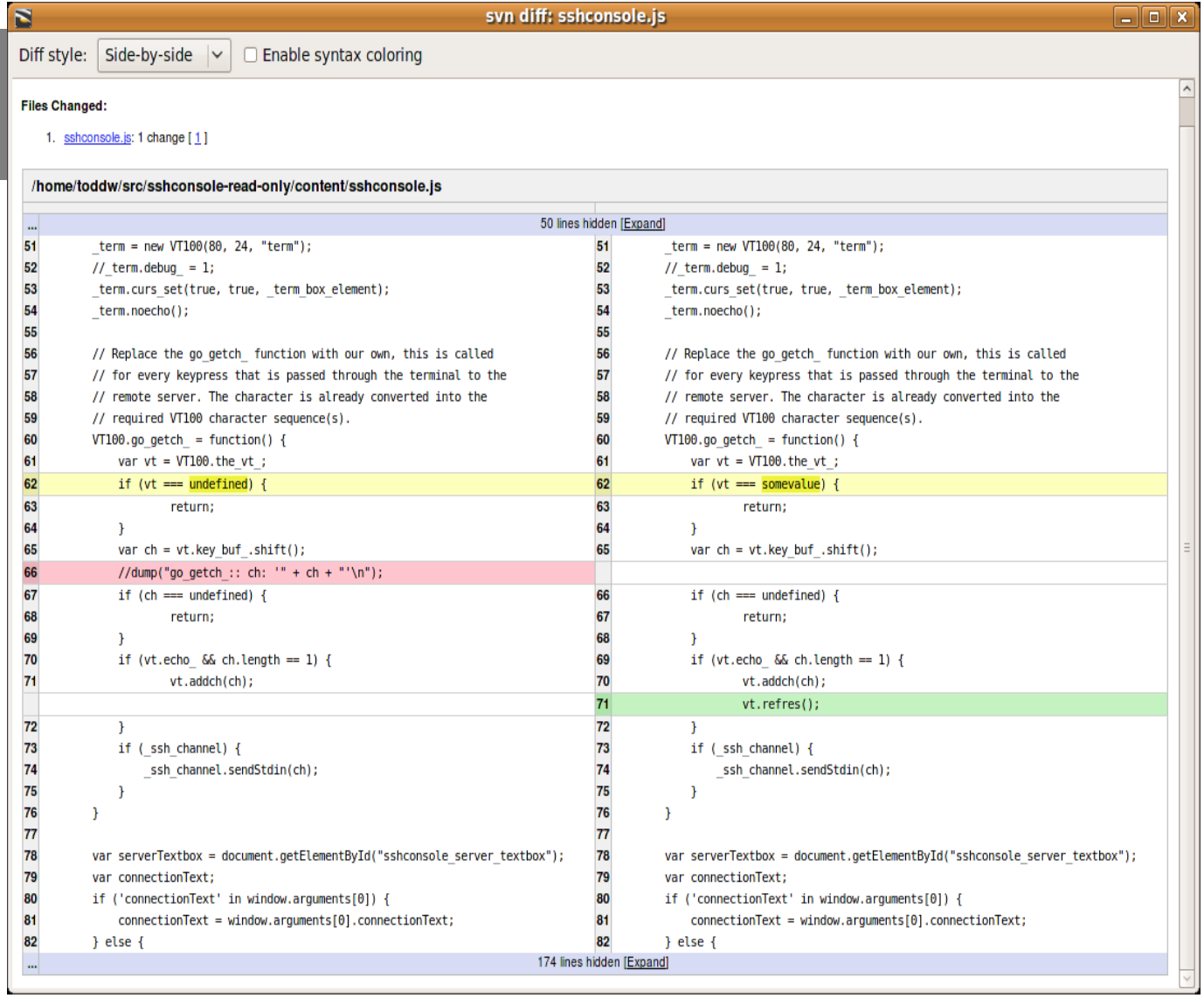
VISUALIZING REVISION HISTORY

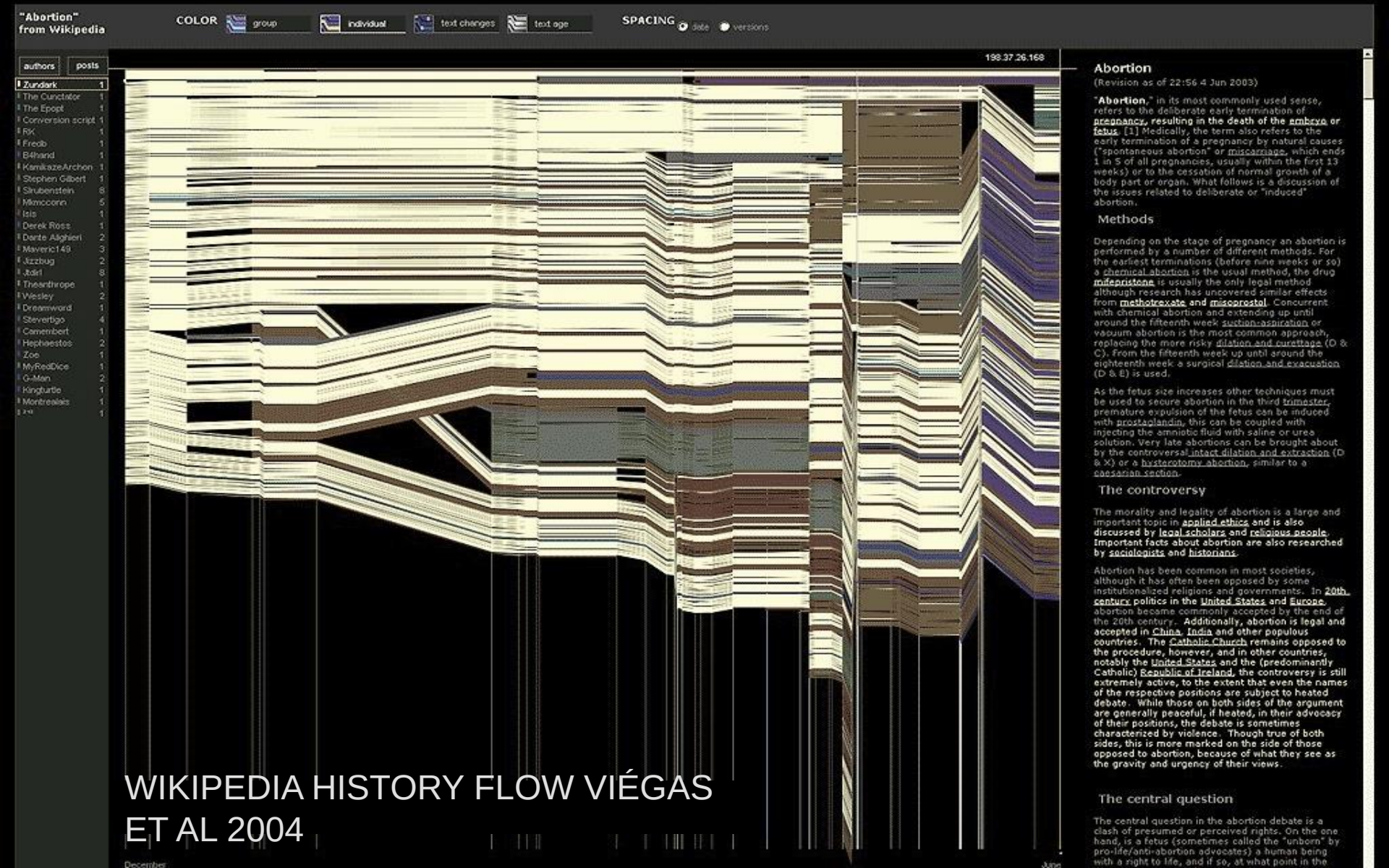
HOW TO DEPICT CONTRIBUTIONS AND CHANGES OVER TIME?





DIFF


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



WIKIPEDIA HISTORY FLOW VIÉGAS ET AL 2004

ANIMATED TRACES

fathom.info/traces

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

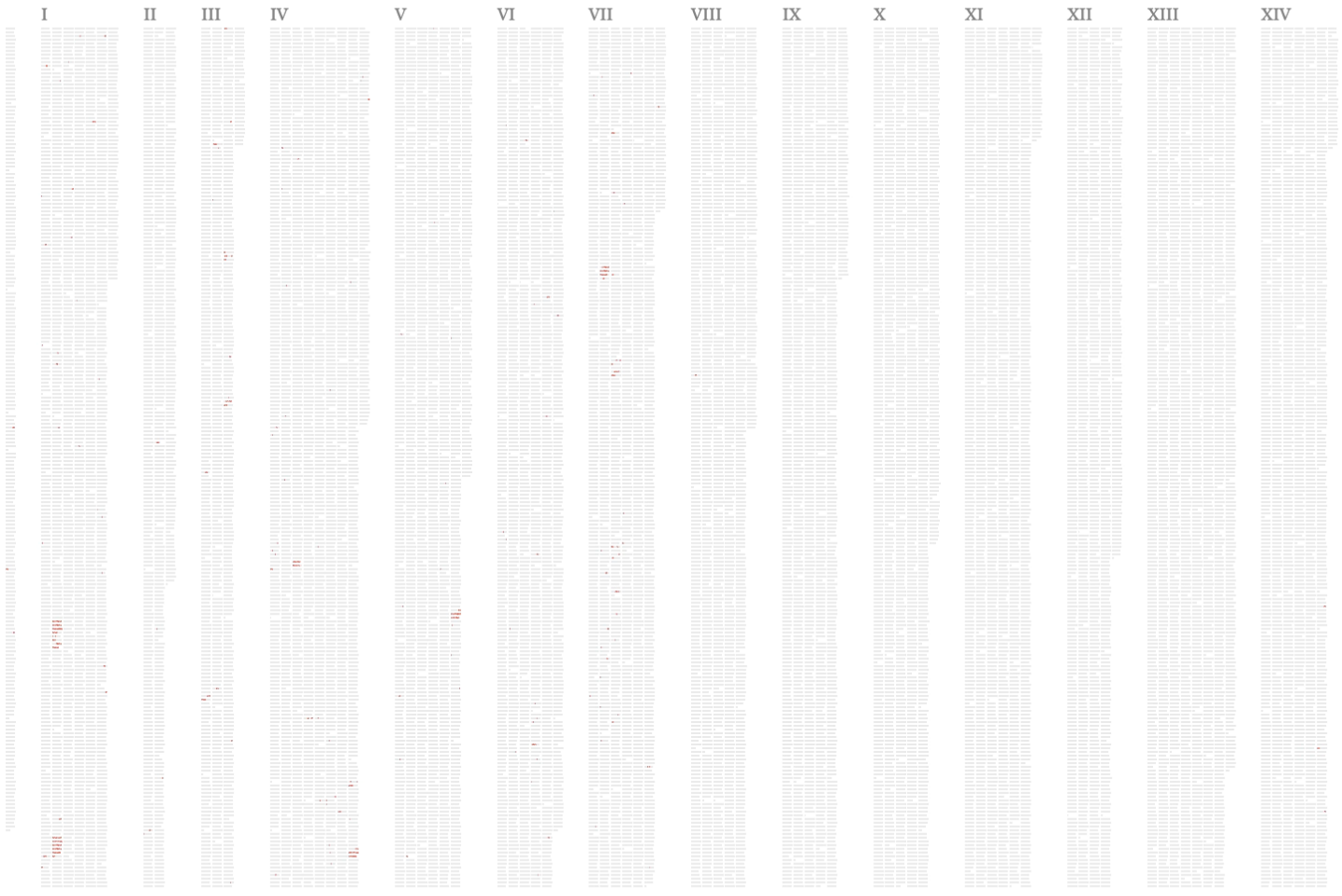
I II III IV V VI VII VIII IX

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 Preserration of Favoured Traces*

Reset Pause Slow Fast



First Edition (1859) Second Edition (1860) Third Edition (1861) Fourth Edition (1866) Fifth Edition (1869) Sixth Edition (1872)

DIFFAMATION

Using Text Animated Transitions to support Navigation in Document Histories

Fanny Chevalier

Pierre Dragicevic

Anastasia Bezerianos

Jean-Daniel Fekete

VISUALIZING DOCUMENT COLLECTIONS

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

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

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

'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

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

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

One dead, another injured, in avalanche near Golden

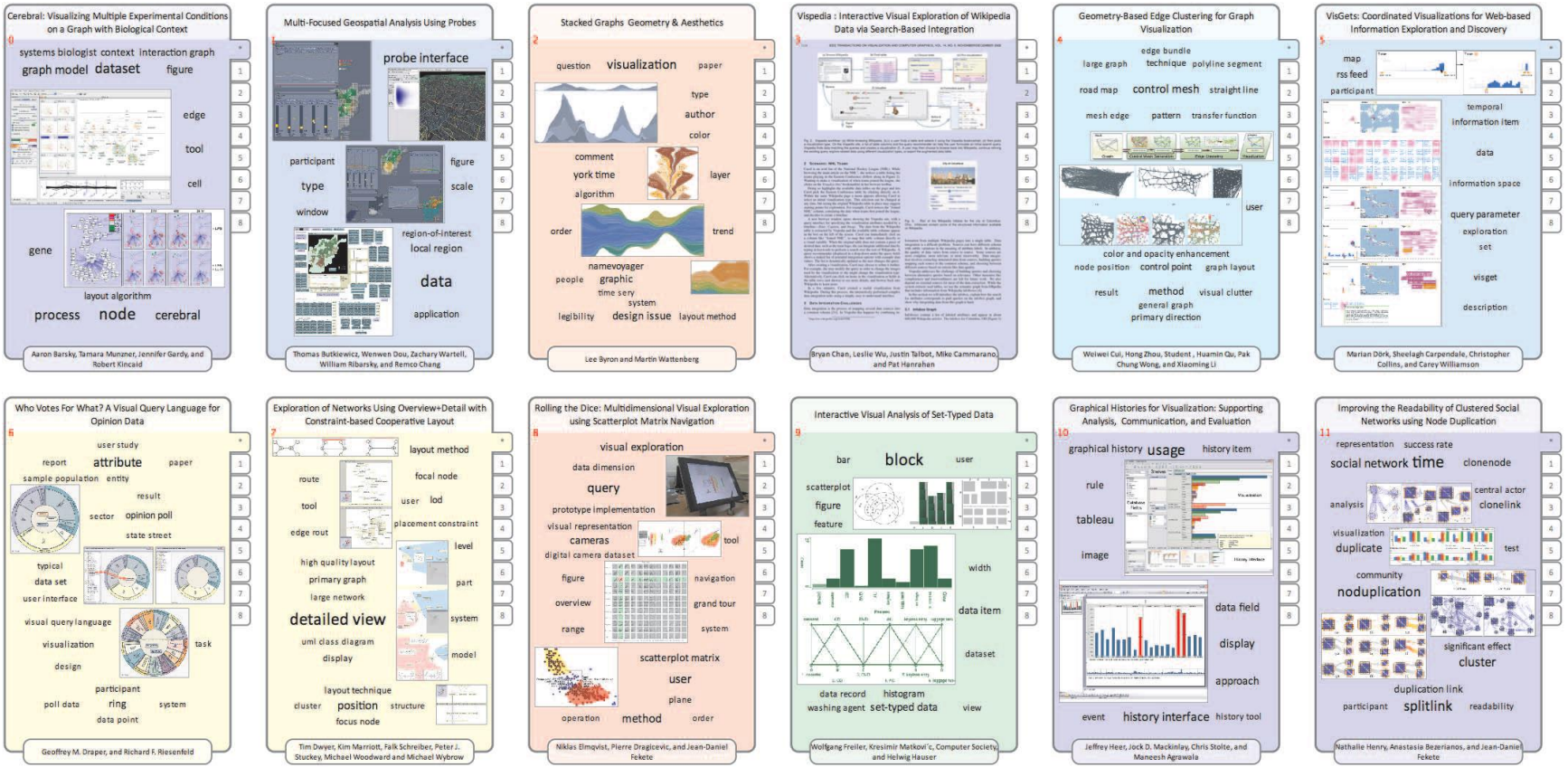
La Loche staff, students return to school this week

HPV cases drop since vaccinations started

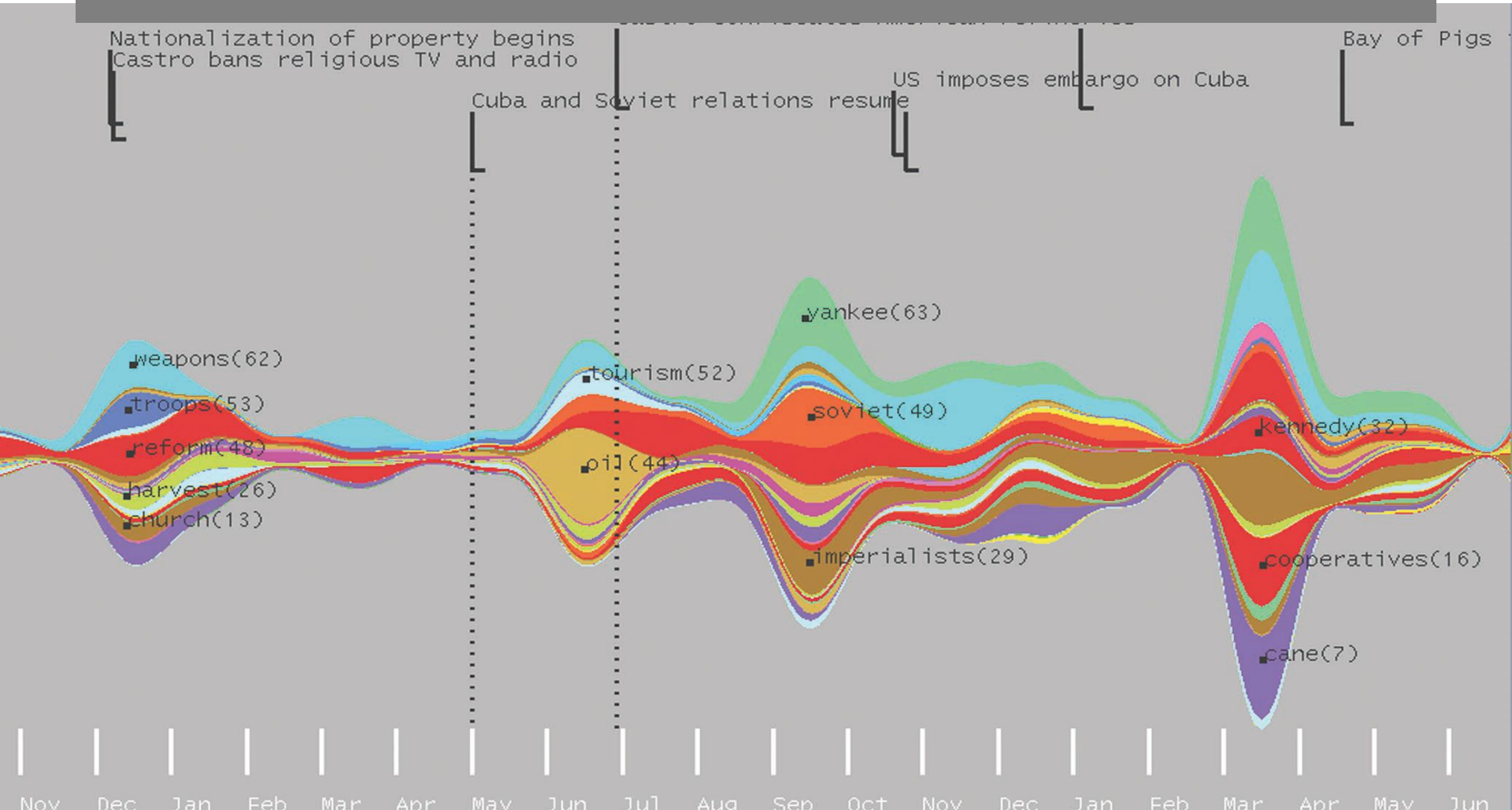
Asian stocks rebound in anticipation of G20 meeting

DOCUMENT CARDS

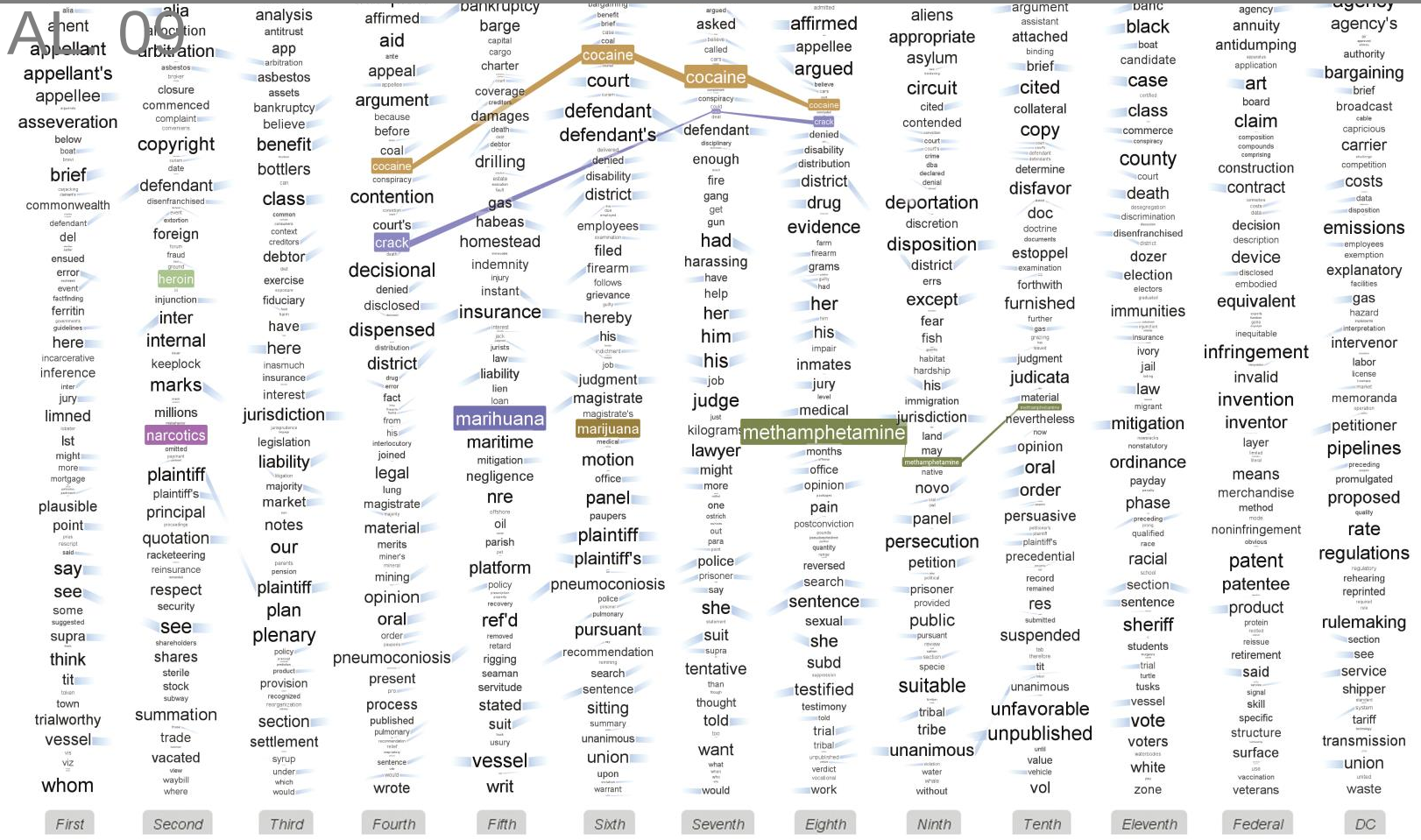
SMALL MULTIPLES FOR DOCUMENTS



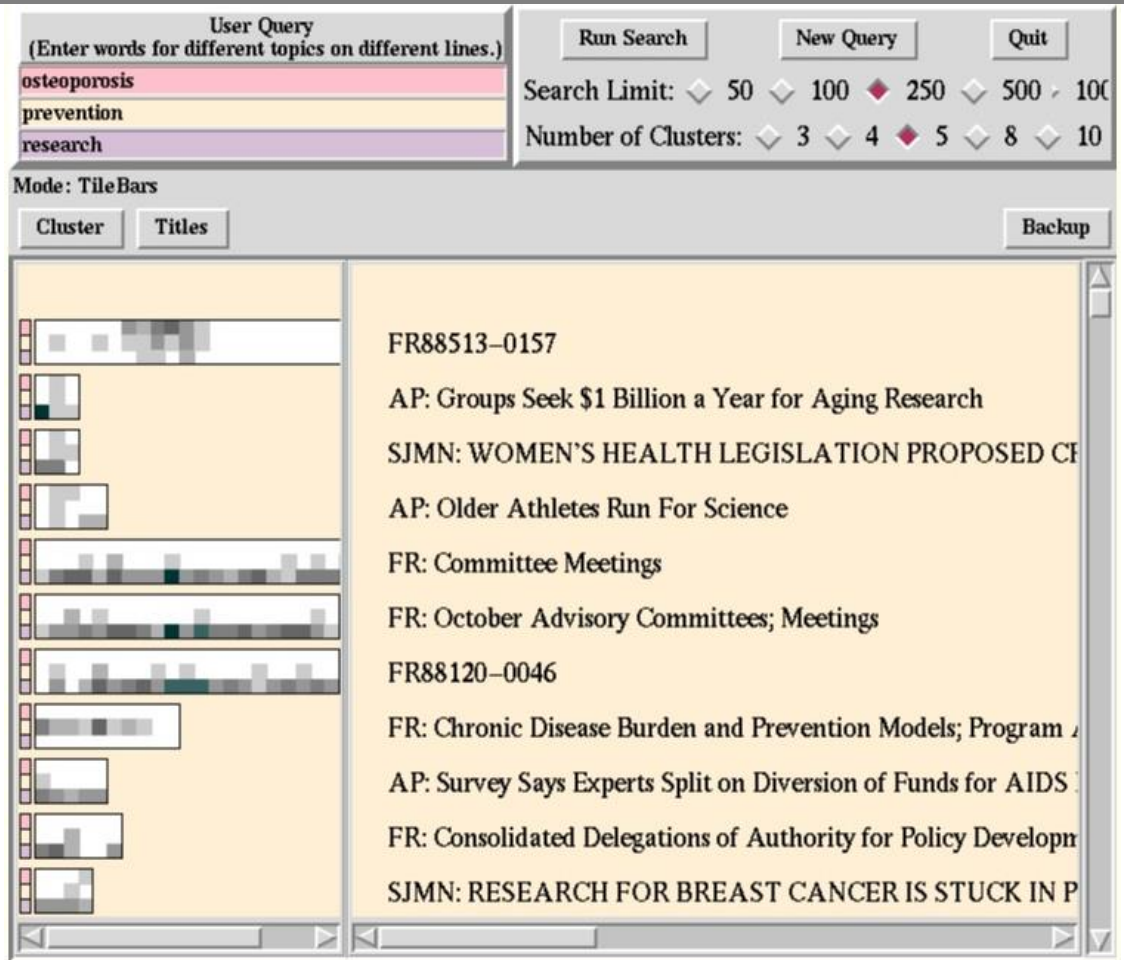
THEMERIVER HAVRE ET AL 1999



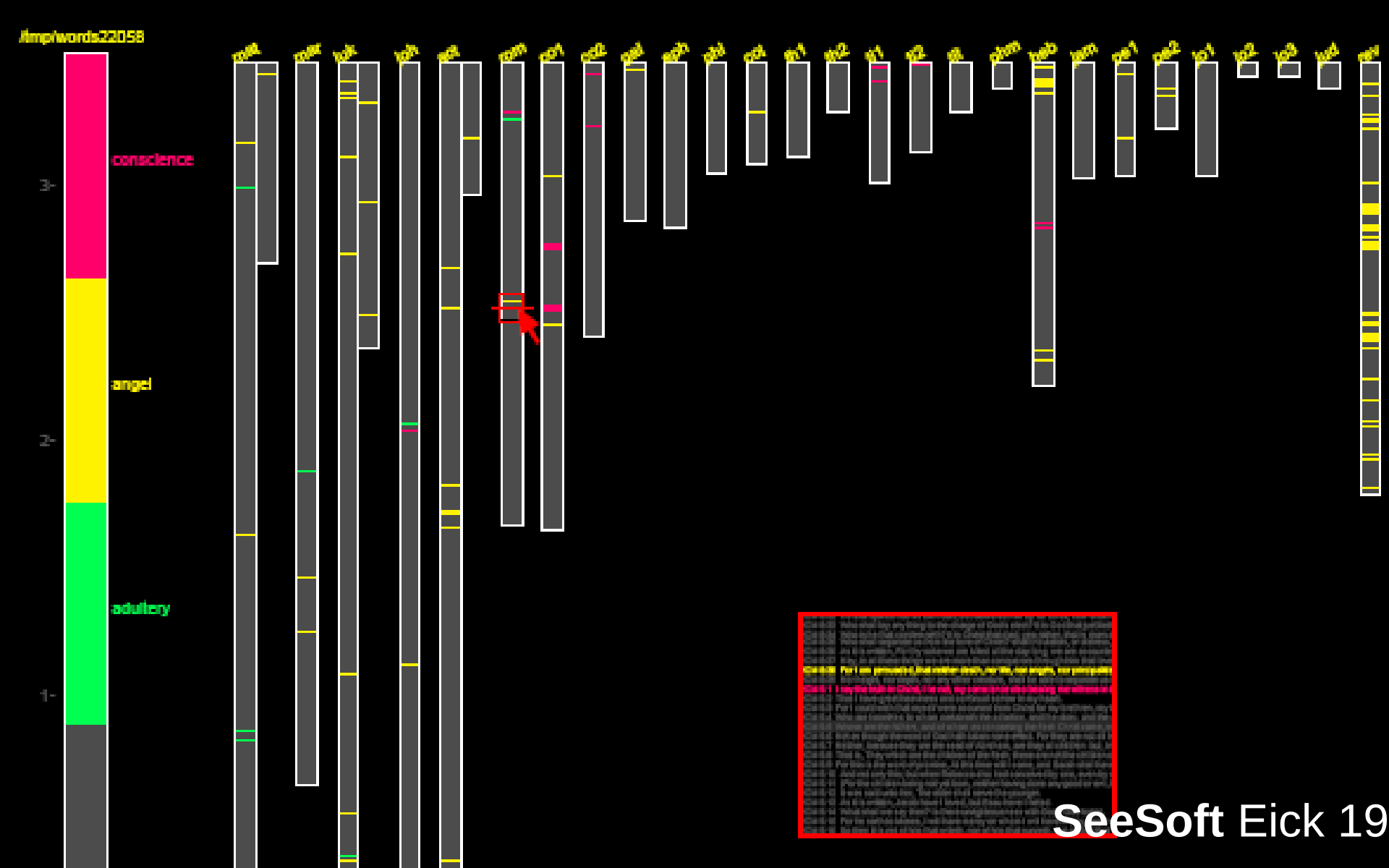
PARALLEL TAG CLOUDS



SUPPORTING SEARCH



TileBars Hearst 1999


/tmp/words22058
conscience
angel
adultery
3
2
1
SeeSoft Eick 19

THE WORDS THAT WERE USED

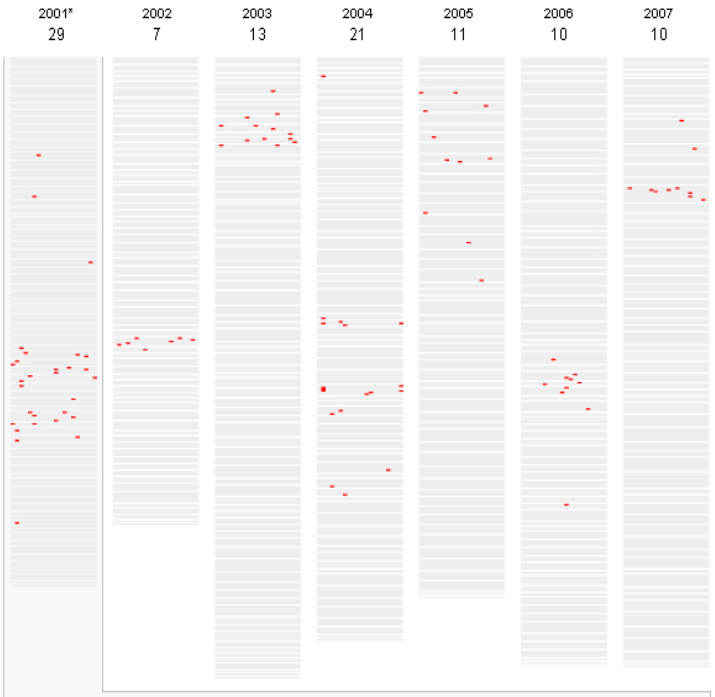
READ 2007 SPEECH | FEEDBACK

The 2007 State of the Union Address

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

Search | or choose a word here.

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



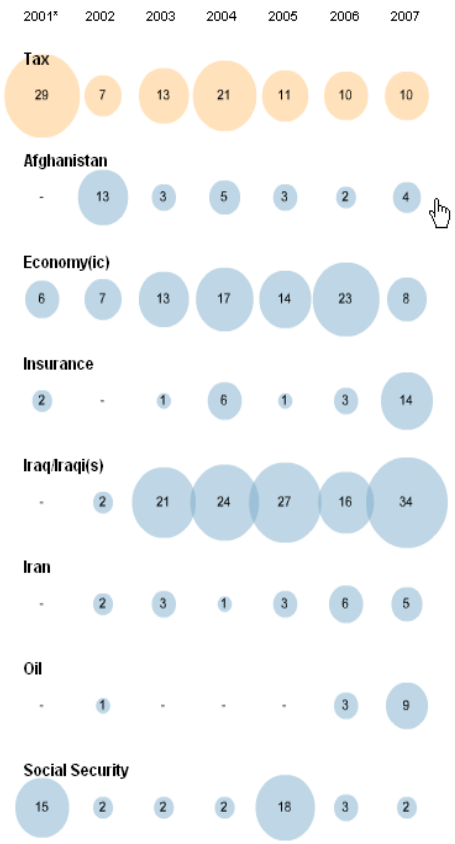
The word in context

Next Instance of 'Tax'

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

-- 2001 (Paragraph 14 of 73)

Compared with other words



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

Ben Werschkul/The New York Times

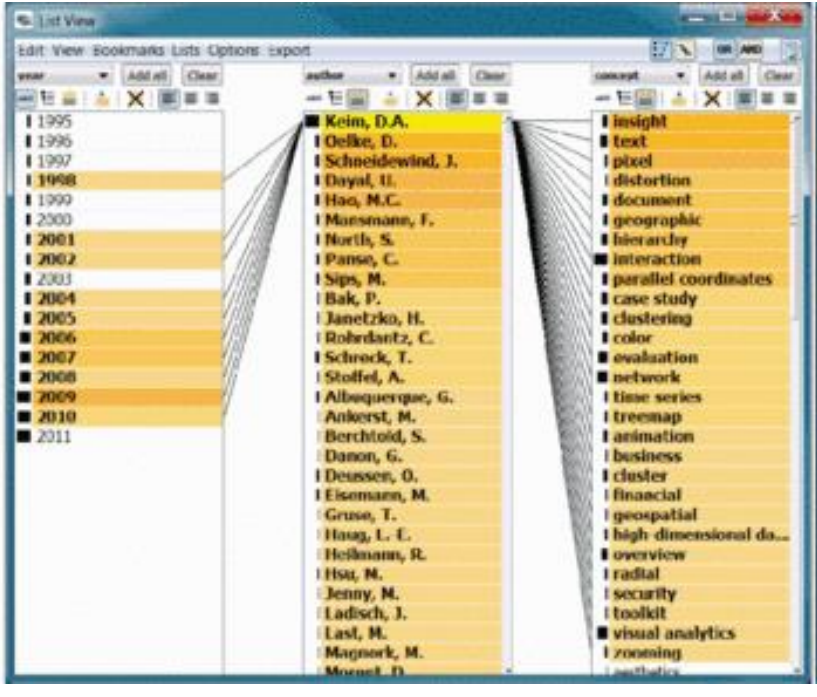
NAMED ENTITY RECOGNITION

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

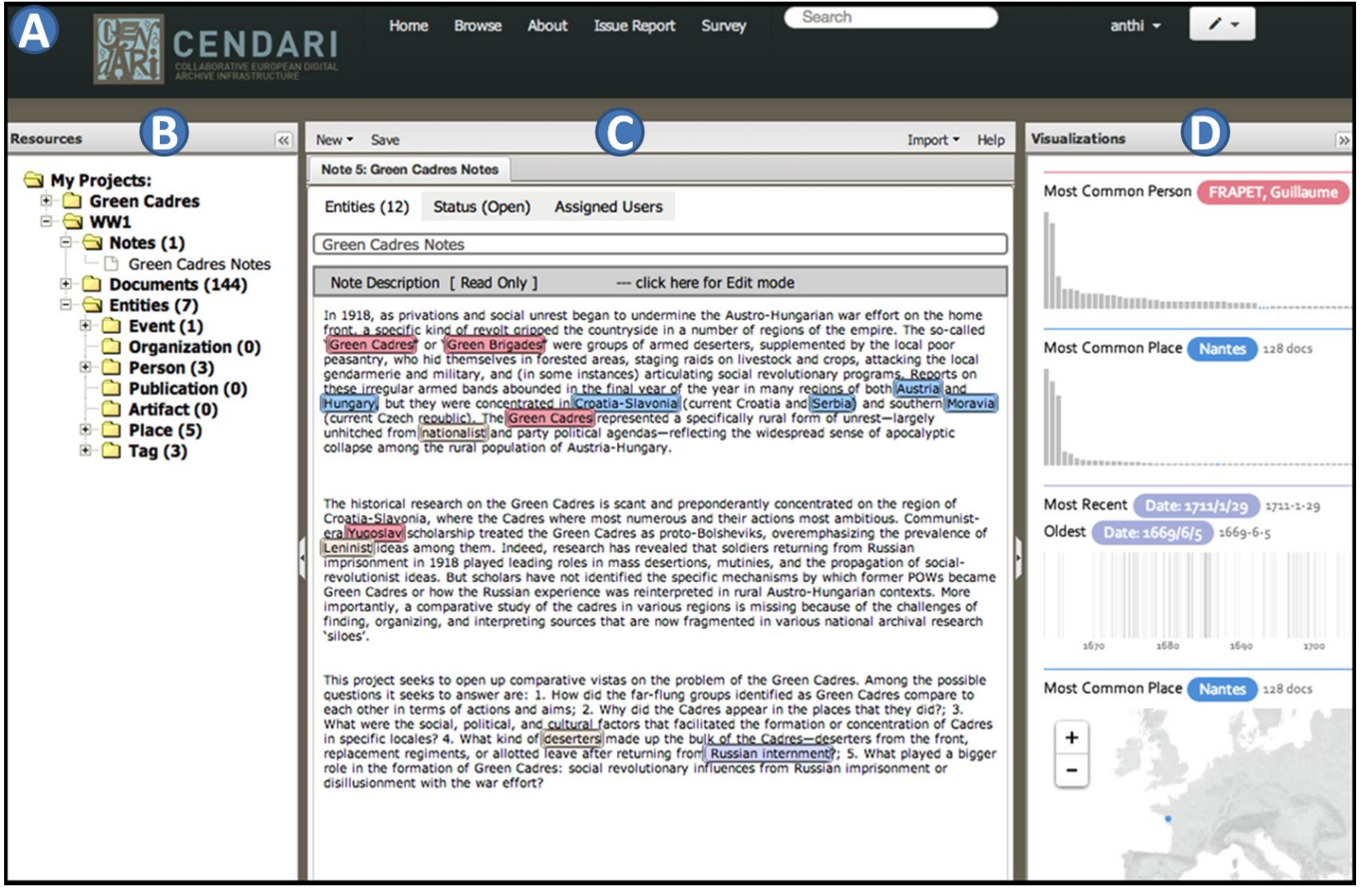
ENTITY RELATIONS: HOW DO THE ENTITIES RELATE?

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


List View
Edit View Bookmarks Lists Options Export
year
author
concept
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.
Haug, L.-E.
Heilmann, R.
Hsu, N.
Jenny, M.
Ladisch, J.
Last, M.
Magnork, M.
Mansmann, F.
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
animation,trends,causality
transform,quality,studied
insight,genes,expression
tables,database,interfaces
classification,geographic,statistics
treemaps,coloring,hierarchically
dimensions,coordinates,parallel
network,graph,social
graphs,edge,algorithm
spaces,internet,search
interact,understand,cognition
collaborative,uses,framework
diverse,environments,toolkit
text,features,topic
video,explorer,stories
querying,series,temporal
3d,displays,navigation
analytics,anomalies,detect
history,mining,patterns
state,displayed,explored

JIGSAW



CENDARI NOTE-TAKING ENVIRONMENT 2015

DOCUMENT SIMILARITY & CLUSTERING

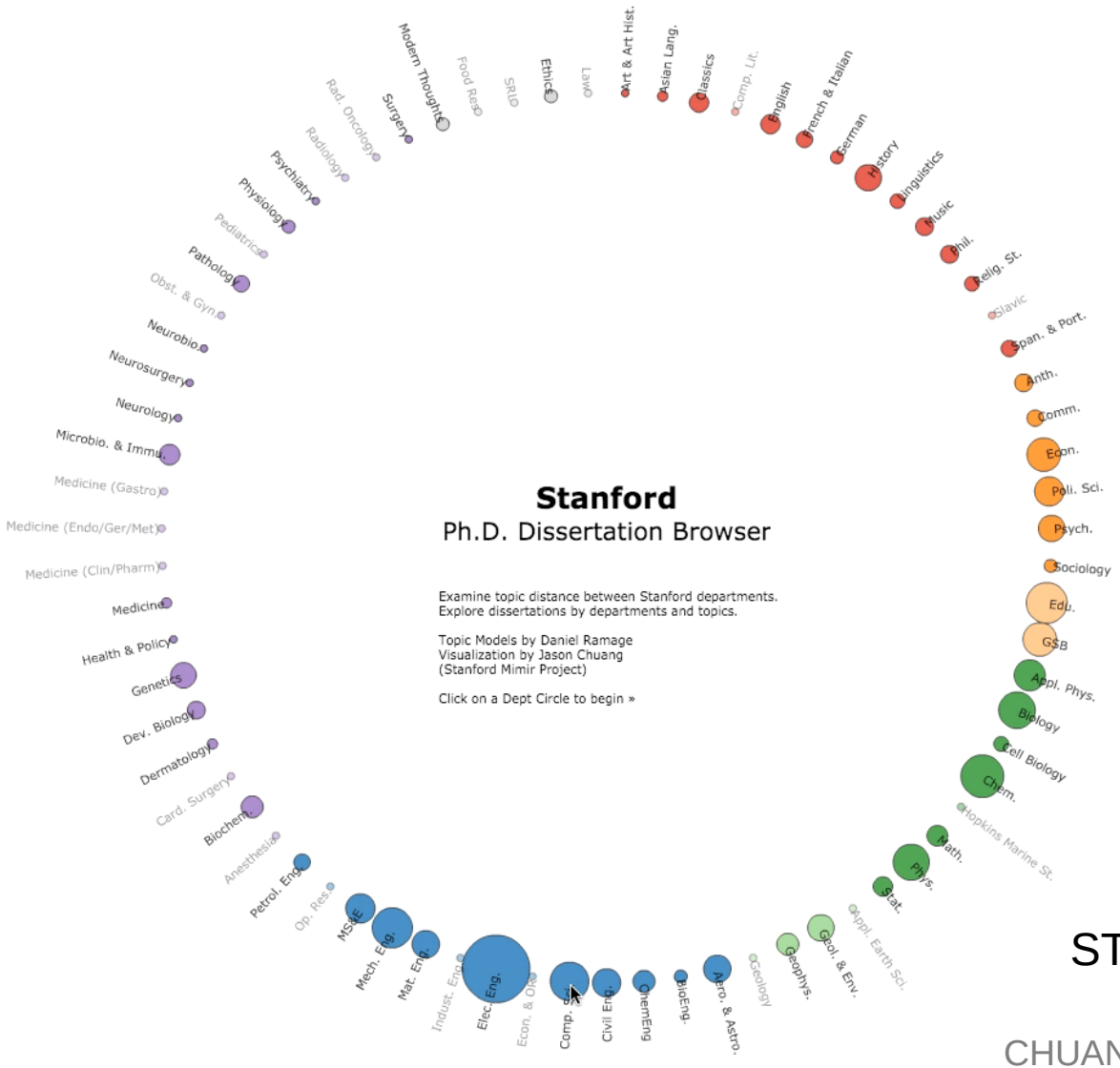
COMPUTE SIMILARITY BETWEEN DOCUMENTS BASED ON THE WORDS THEY SHARE

- 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

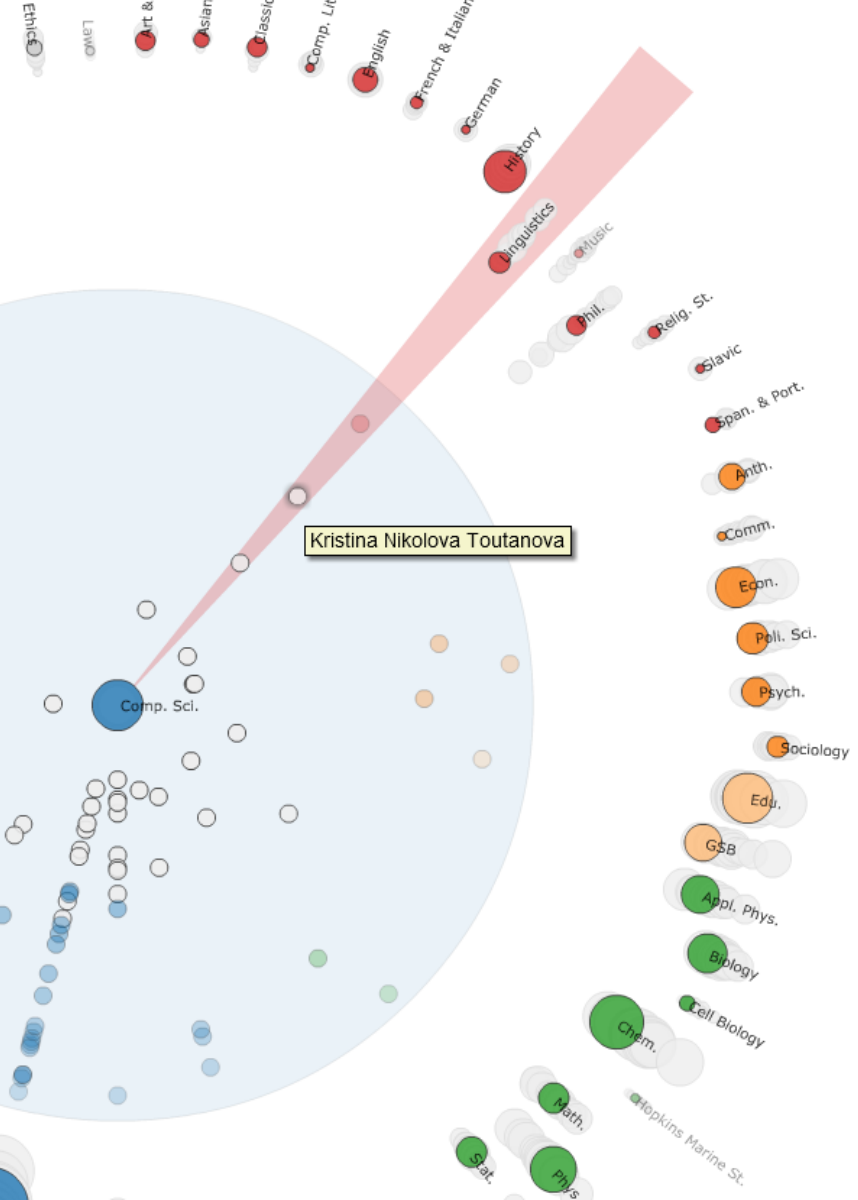
- MANY, MANY APPROACHES EXIST



STANFORD 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.

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)

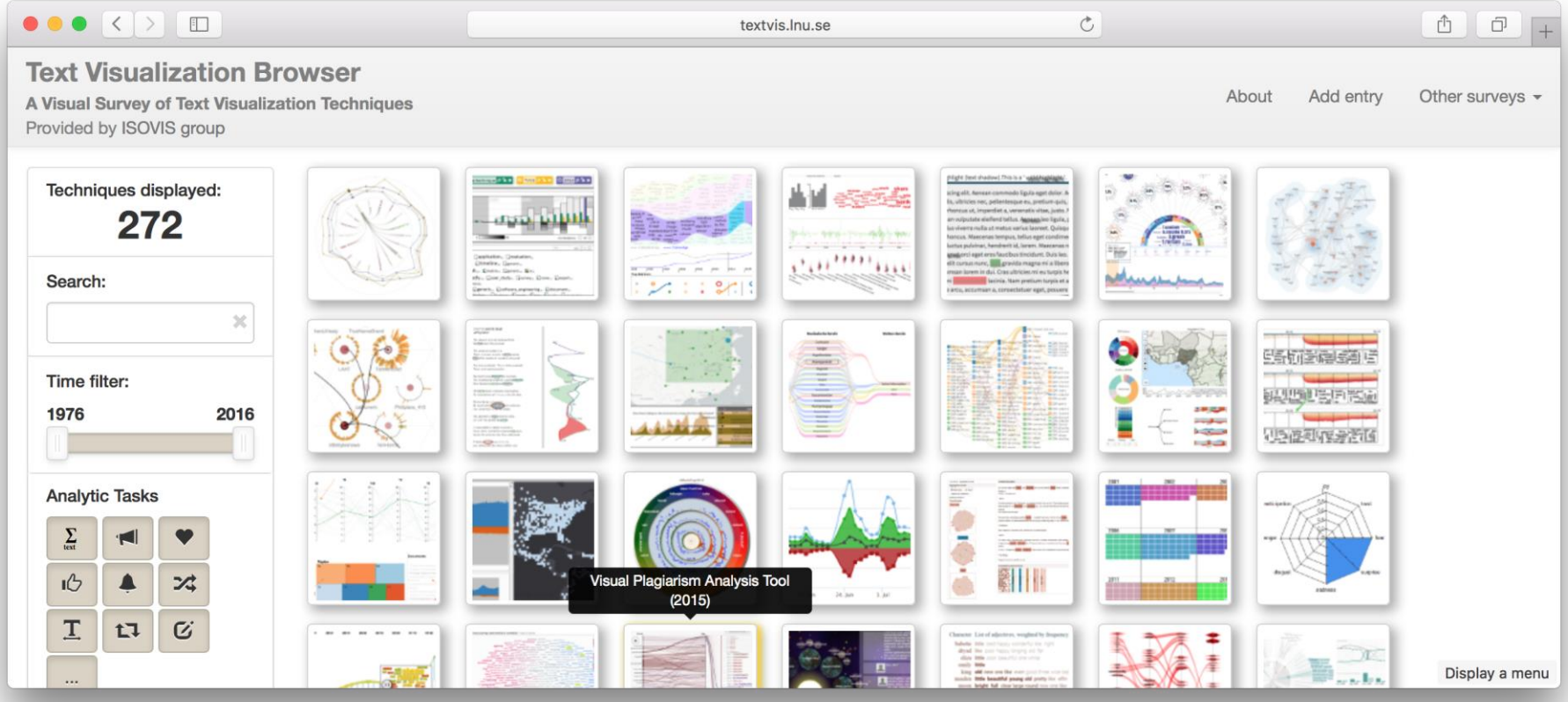
WHERE POSSIBLE, USE VISUALIZATION AS INDEX INTO DOCUMENTS

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/



QUESTIONS?

ACKNOWLEDGEMENTS

Slides in were inspired, adapted, taken from slides by

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