

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



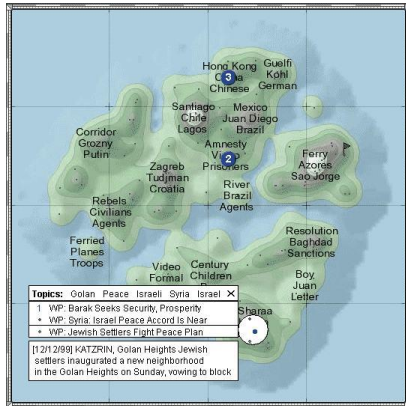
(TODAY)

VISUALIZING TEXT

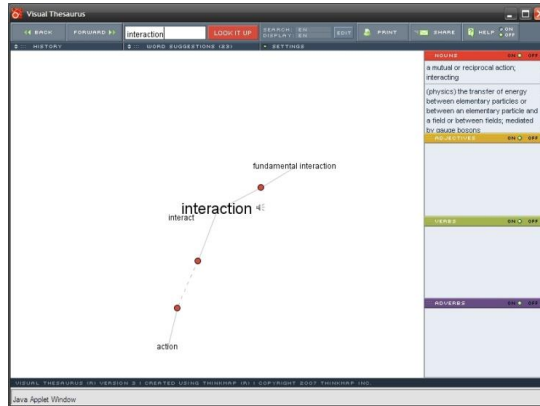
TEXT?

WHY

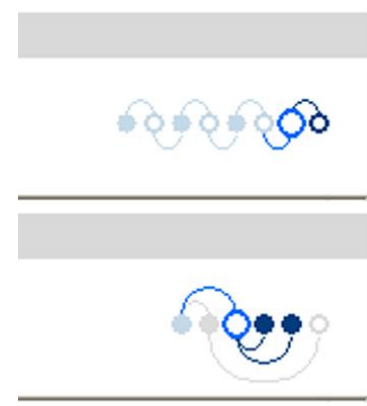
- To assist information retrieval
- To enable linguistic analysis
- To augment analytics on mixed data



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

- Millions of blog posts,
- Hundreds of thousands of news stories,
- 183 billion emails,
- ... **per day**

NOISY DATA

- 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:
 - You are a couch potato.
 - They fought like cats and dogs.
 - Opportunity knocked on the door

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...

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 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)

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?

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)

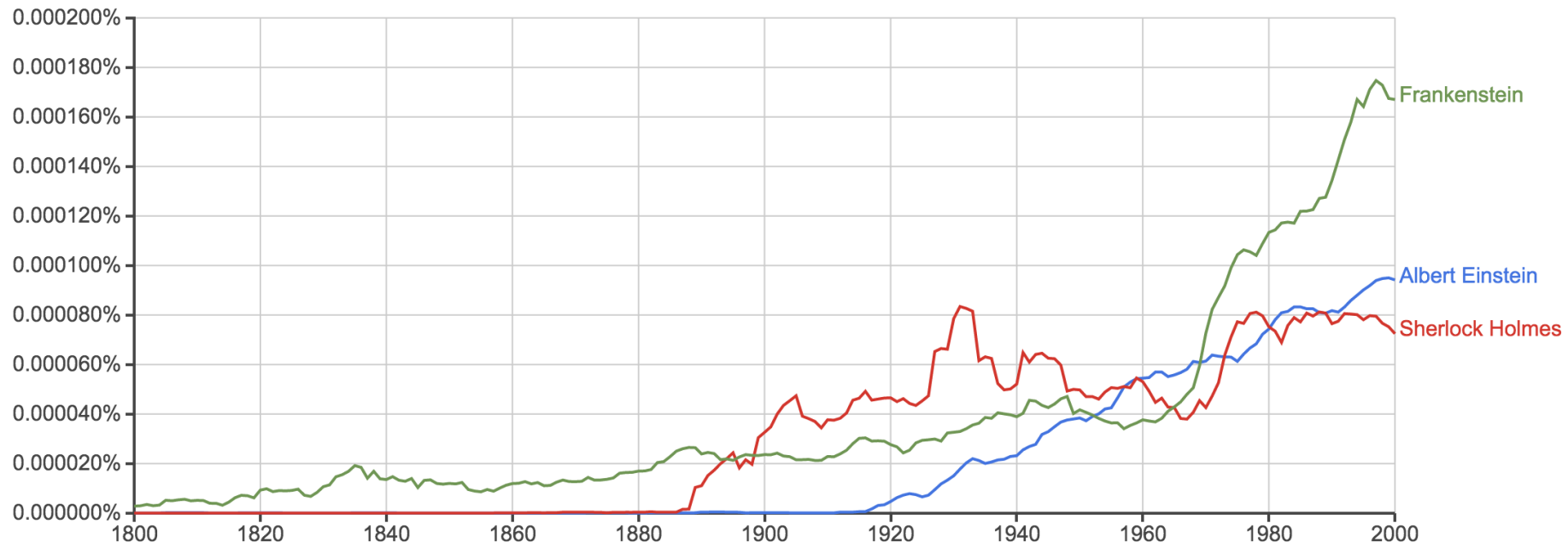
[the quick brown fox jumps], [quick brown fox jumps over], [brown fox jumps over]

...

Google Books Ngram Viewer

Graph these comma-separated phrases: case-insensitive

between and from the corpus with smoothing of [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

TAG CLOUDS

WORD COUNT

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 oae
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**

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 oae
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

WORDCOUNT

WORDCOUNT

◀ PREVIOUS WORD

NEXT WORD ▶

the of and to in that it is was i for on you he be with as by a have are this no but had his they from she which we in there were do you it is has you will for when re ce who about per said the con the

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50

CURRENT WORD

FIND WORD:

BY RANK:

REQUESTED WORD: THE

RANK: 1

ARCHIVE

COUNT

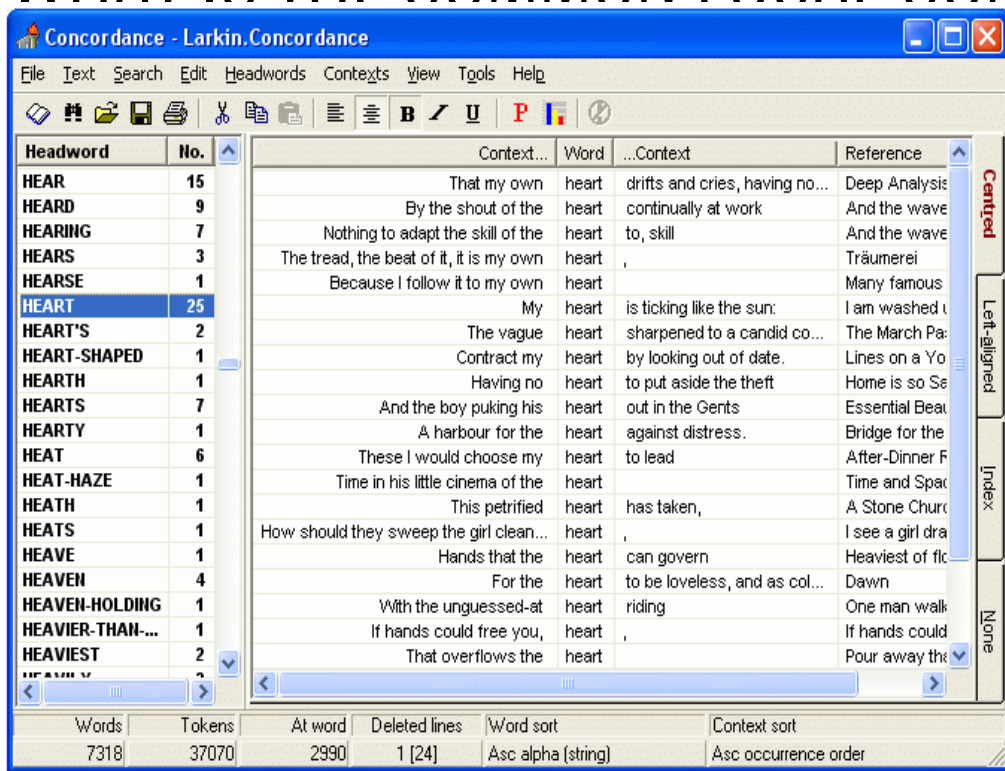


JONATHAN HARRIS

<http://wordcount.org>

CONCORDANCE

WHAT IS THE COMMON LOCAL CONTEXT OF A TERM?



The screenshot shows the Larkin Concordance software interface. The main window displays a list of words and their occurrences in various contexts. The word 'HEART' is highlighted in blue, indicating it is the current selection. The interface includes a menu bar (File, Text, Search, Edit, Headwords, Contexts, View, Tools, Help), a toolbar with icons for file operations and editing, and a status bar at the bottom showing statistics: Words (7318), Tokens (37070), At word (2990), Deleted lines (1 [24]), Word sort (Asc alpha (string)), and Context sort (Asc occurrence order).

Headword	No.	Context...	Word	...Context	Reference
HEAR	15		heart	drifts and cries, having no...	Deep Analysis
HEARD	9	By the shout of the	heart	continually at work	And the wave
HEARING	7	Nothing to adapt the skill of the	heart	to, skill	And the wave
HEARS	3	The tread, the beat of it, it is my own	heart	,	Träumerei
HEARSE	1	Because I follow it to my own	heart		Many famous
HEART	25	My	heart	is ticking like the sun:	I am washed t
HEART'S	2	The vague	heart	sharpened to a candid co...	The March Pa
HEART-SHAPED	1	Contract my	heart	by looking out of date.	Lines on a Yo
HEARTH	1	Having no	heart	to put aside the theft	Home is so Se
HEARTS	7	And the boy puking his	heart	out in the Gents	Essential Bea
HEARTY	1	A harbour for the	heart	against distress.	Bridge for the
HEAT	6	These I would choose my	heart	to lead	After-Dinner F
HEAT-HAZE	1	Time in his little cinema of the	heart		Time and Spa
HEATH	1	This petrified	heart	has taken,	A Stone Churr
HEATS	1	How should they sweep the girl clean...	heart	,	I see a girl dra
HEAVE	1	Hands that the	heart	can govern	Heaviest of flc
HEAVEN	4	For the	heart	to be loveless, and as col...	Dawn
HEAVEN-HOLDING	1	With the unguessed-at	heart	riding	One man walk
HEAVIER-THAN...	1	If hands could free you,	heart	,	If hands could
HEAVIEST	2	That overflows the	heart		Pour away th

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



love the

lord

thy god

with all

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

and with all thy might .
 that thou mayest live .

mind
 strength , a

and

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

and to

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

your god

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

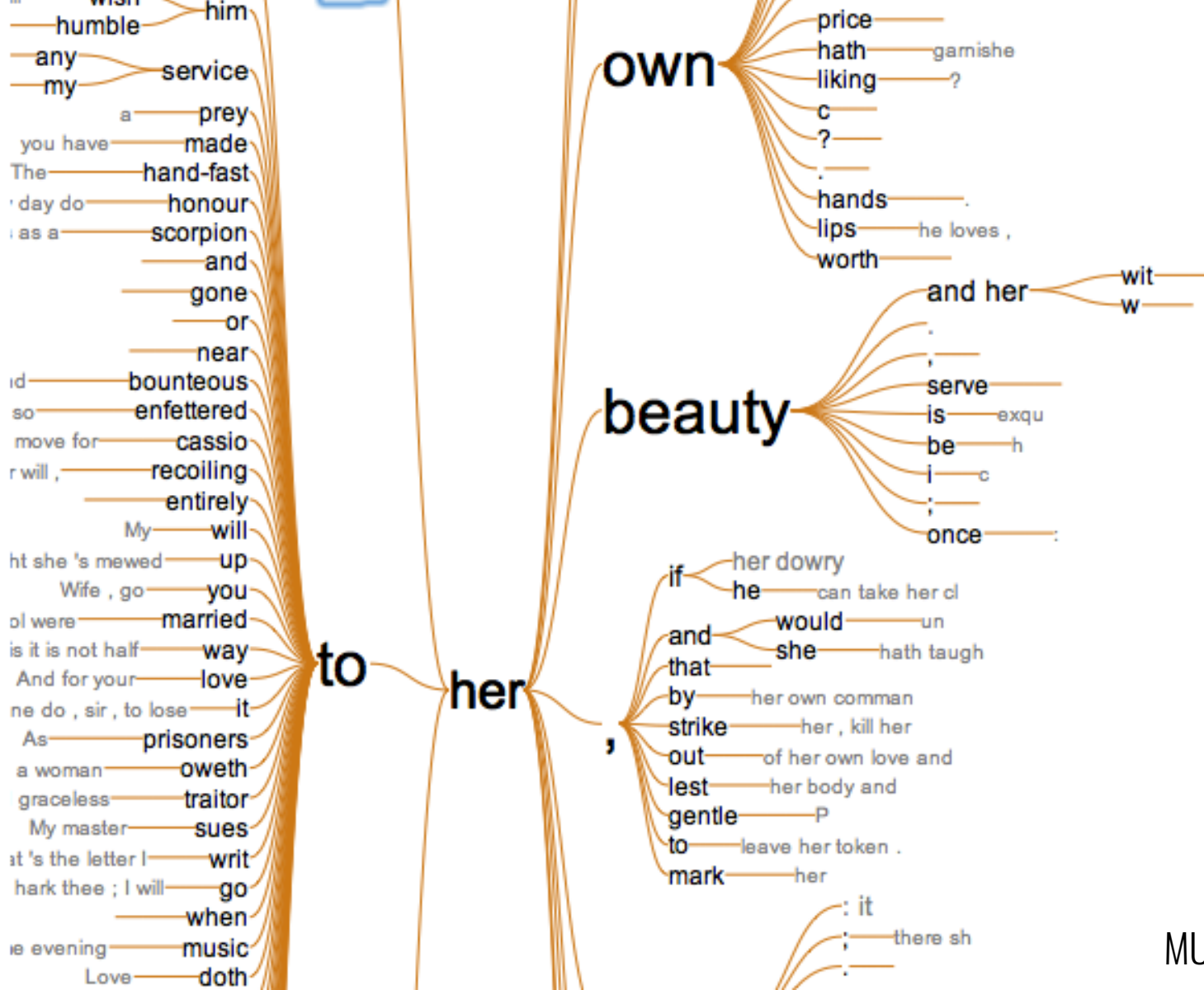
uppermost

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 .

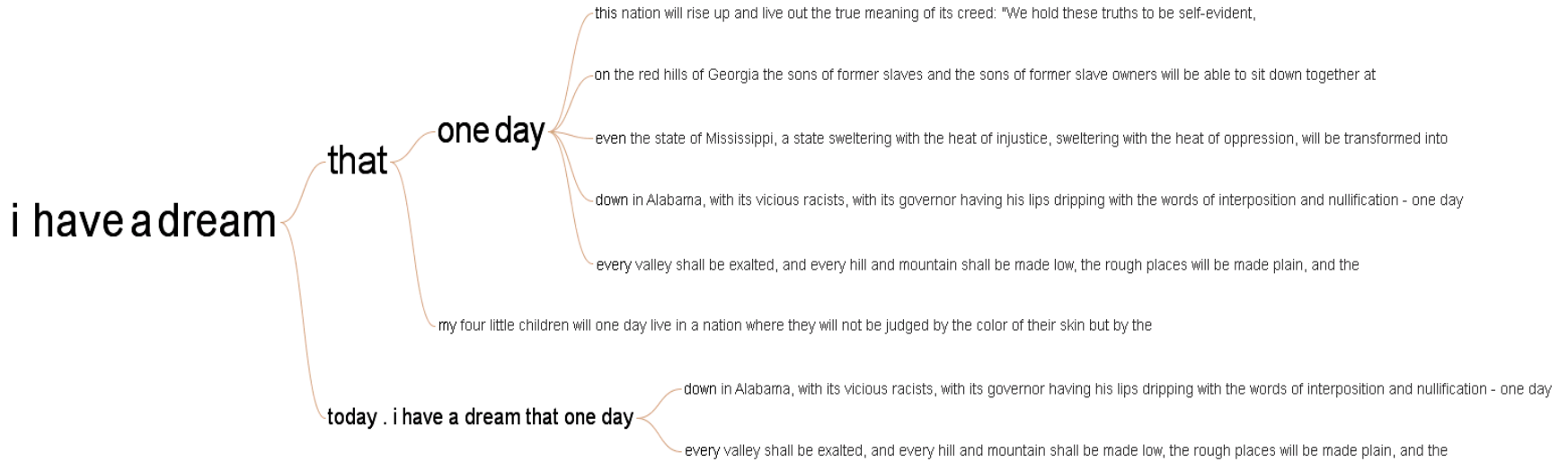
father

; 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 .



RECURRENT THEMES IN SPEECH



GLIMPSES OF STRUCTURE

CONCORDANCES SHOW LOCAL, REPEATED
STRUCTURE

BUT WHAT ABOUT OTHER TYPES OF PATTERNS?

FOR EXAMPLE

LEXICAL: <A> at

SYNTACTIC: <Noun> <Verb> <Object>

PHRASE NETS

LOOK FOR SPECIFIC LINKING PATTERNS IN THE TEXT:

'A **AND** B', 'A **AT** B', 'A **OF** B', ETC

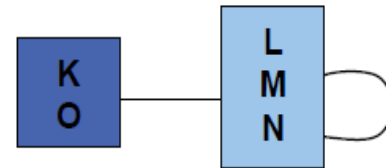
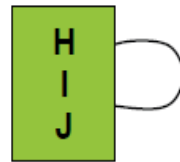
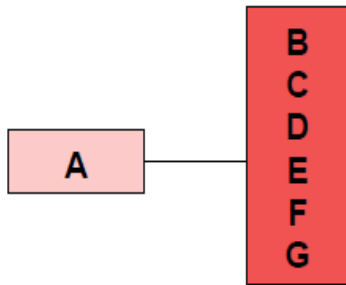
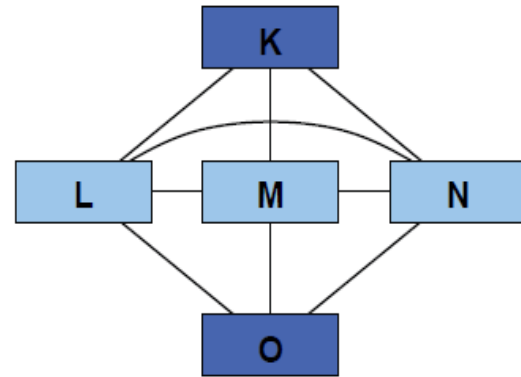
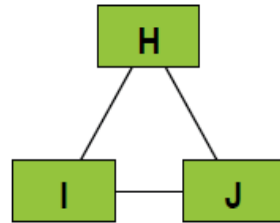
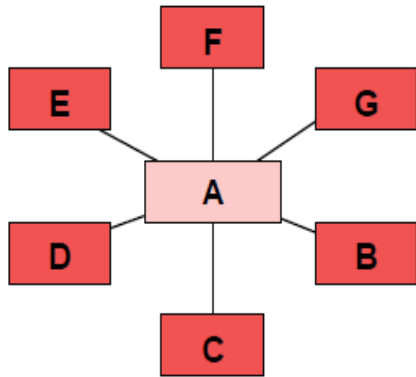
COULD BE OUTPUT OF REGEXP OR PARSER

VISUALIZE EXTRACTED PATTERNS IN A NODE-LINK VIEW

OCCURRENCES = NODE SIZE

PATTERN POSITION = EDGE DIRECTION

NODE GROUPING



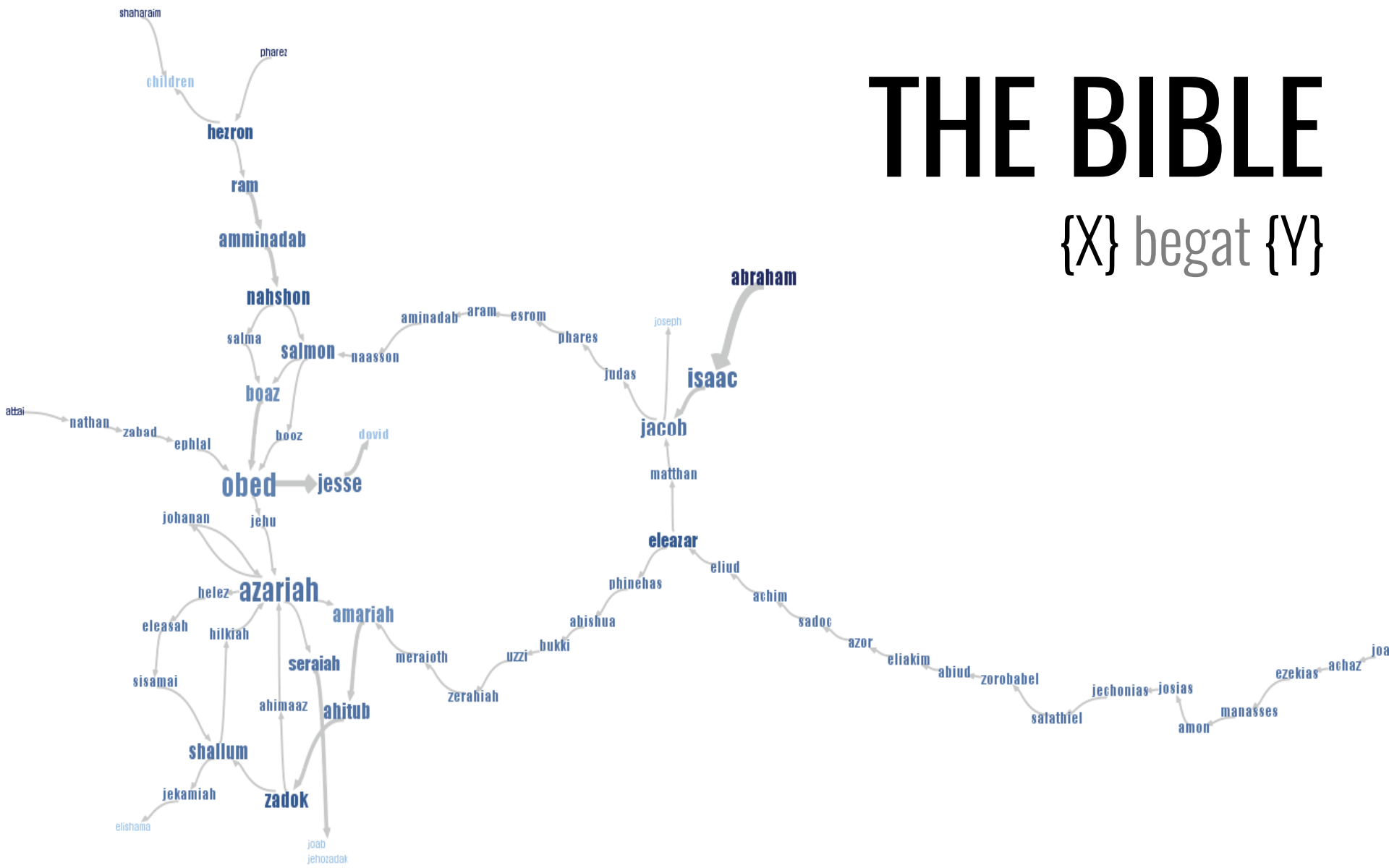
(a)

(b)

(c)

THE BIBLE

{X} begat {Y}



VISUALIZING DOCUMENT COLLECTIONS

Analysts: GOP may regret 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

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

US rejected North Korea peace talks offer before last nuclear test

Malaysia, south-east Asia nations wamed of terror attacks

Samsung, LG unveil new devices in bid for smartphone recovery

Raceline Radio Program Guide: February 21, 2016

Canada, USA renew rivalry in CONCACAF final

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

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

Chan wins Four Continents figure-skating championship

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

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

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

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 Open Folder, the One With a Budget That Talks

Truex comes up a few inches short in closest Daytona 500

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

HPV cases drop since vaccinations started

Asian stocks rebound in anticipation of G20 meeting

Asian stocks rebound in anticipation of G20 meeting

DOCUMENT CARDS

SMALL MULTIPLES FOR DOCUMENTS

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

4 systems biologist context interaction graph graph model dataset figure

1 2 3 4 5 6 7 8

edge tool cell gene layout algorithm process node cerebral

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

Multi-Focused Geospatial Analysis Using Probes

probe interface

1 2 3 4 5 6 7 8

participant type window region-of-interest local region data application

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

Stacked Graphs: Geometry & Aesthetics

question visualization paper

1 2 3 4 5 6 7 8

comment york time algorithm trend nameviewer graphic time sery system legibility design issue layout method

Lee Byron and Martin Wattenberg

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

1 2 3 4 5 6 7 8

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

Geometry-Based Edge Clustering for Graph Visualization

4 edge bundle large graph road map mesh edge color and opacity enhancement node position result

1 2 3 4 5 6 7 8

technique polyline segment control mesh transfer function pattern user control point graph layout method visual cluster general graph primary direction

Weiwei Cui, Hong Zhou, Student1, Huamin Qu, Pak Chung Wong, and Xiaoming Li

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

5 map rss feed participant

1 2 3 4 5 6 7 8

temporal information item data information space query parameter exploration set visget description

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

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

6 user study report attribute paper sample population entity result sector opinion poll state street typical data set user interface visual query language visualization design participant poll data ring position data point

1 2 3 4 5 6 7 8

task

Geoffrey M. Drapes, and Richard F. Riesenfeld

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

7 layout method route tool edge route high quality layout primary graph large network detailed view uml class diagram display layout technique cluster position focus node

1 2 3 4 5 6 7 8

focal node user lod placement constraint level part system model structure

Tim Dawyer, Kim Marriott, Falk Schreiber, Peter J. Stuckey, Michael Woodward and Michael Wybro

Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation

8 visual exploration data dimension query prototype implementation visual representation cameras digital camera dataset figure overview range scatterplot matrix user operation method order

1 2 3 4 5 6 7 8

navigation grand tour system plane

Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete

Interactive Visual Analysis of Set-Typed Data

9 bar block user scatterplot figure feature width data item dataset data record washing agent set-typed data view

1 2 3 4 5 6 7 8

Wolfgang Freiler, Kresimir Matkovic, Computer Society, and Helwig Hauser

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

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

1 2 3 4 5 6 7 8

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

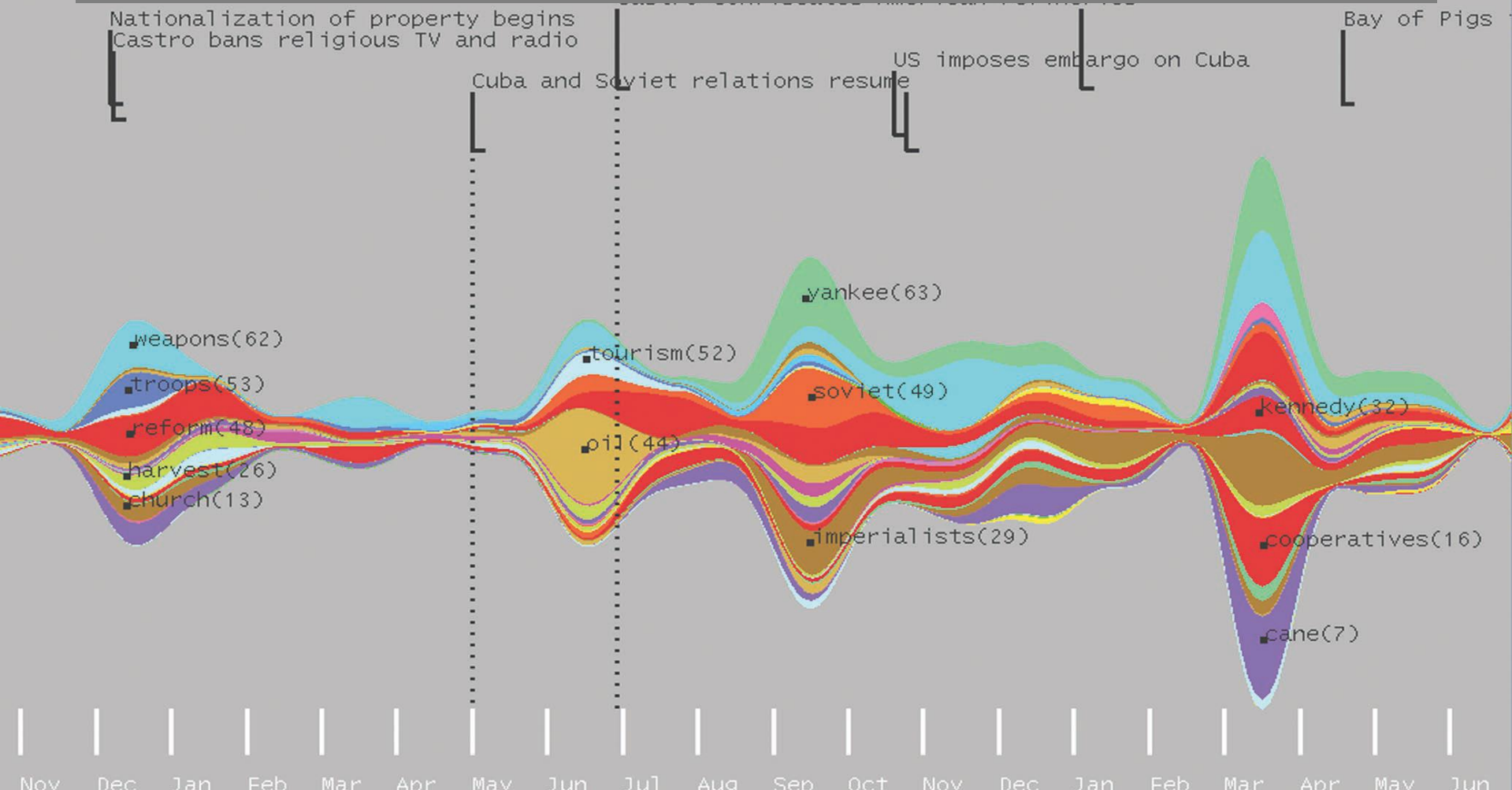
Improving the Readability of Clustered Social Networks using Node Duplication

11 representation success rate social network time clonemode analysis visualization duplicate community noduplication visualization significant effect cluster duplication link participant splitlink readability

1 2 3 4 5 6 7 8

Nathalie Henry, Anastasia Bezerianos, and Jean-Daniel Fekete

THEMERIVER HAVRE ET AL 1999



SUPPORTING SEARCH

The screenshot displays a search interface with the following components:

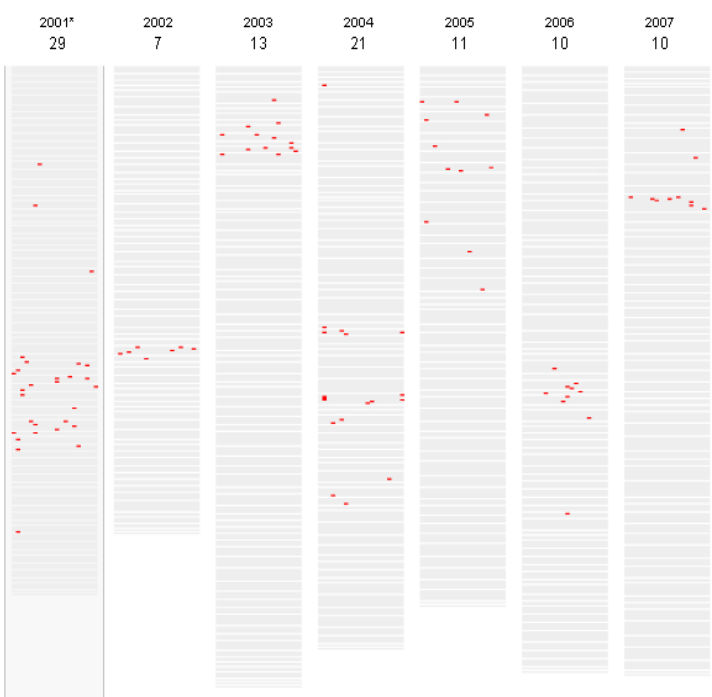
- User Query:** A text area containing "osteoporosis", "prevention", and "research" on separate lines.
- Search Controls:** Buttons for "Run Search", "New Query", and "Quit".
- Search Parameters:** "Search Limit" with a dropdown menu showing 50, 100, 250 (selected), 500, and 1000; and "Number of Clusters" with a dropdown menu showing 3, 4, 5 (selected), 8, and 10.
- Mode:** A dropdown menu set to "Tile Bars".
- View Options:** Buttons for "Cluster" and "Titles".
- Backup:** A "Backup" button.
- Results:** A list of search results, each with a small bar chart on the left and a text snippet on the right. The results include:
 - 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

TileBars Hearst 1999

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



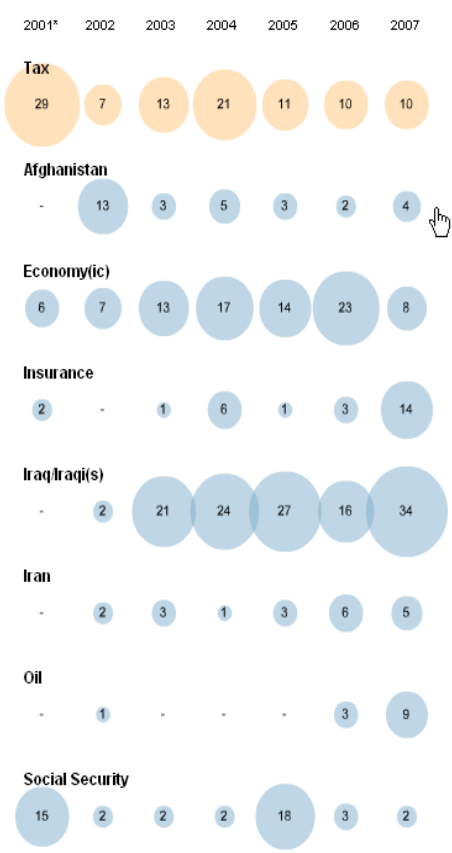
The word in context

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.

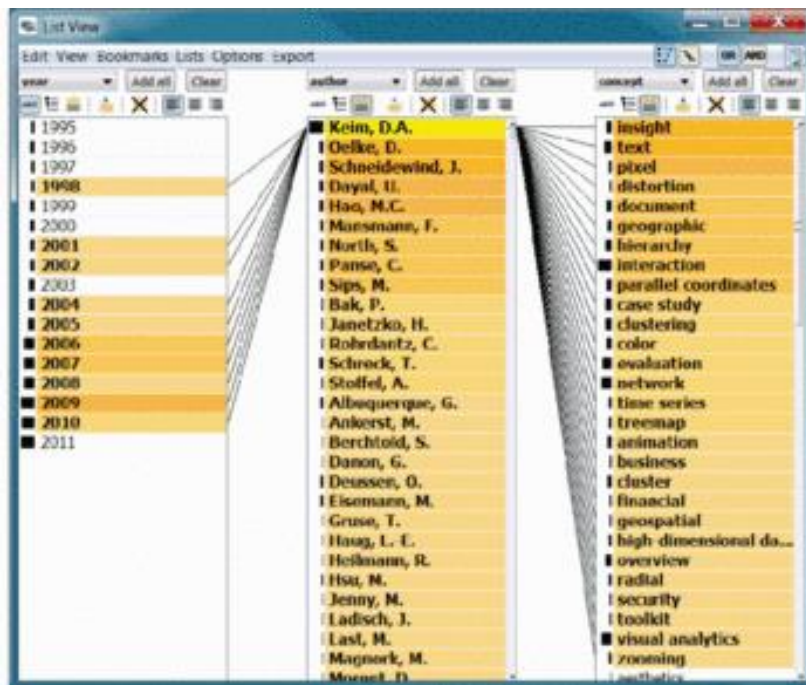
-- 2001 (Paragraph 14 of 73)

[Next instance of 'Tax'](#)


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.



A



CENDARI
COLLABORATIVE EUROPEAN DIGITAL
ARCHIVE INFRASTRUCTURE

Home Browse About Issue Report Survey

Search

anthi

B Resources
C New Save Import Help
D Visualizations

My Projects:

- Green Cadres
- WW1
 - Notes (1)
 - Green Cadres Notes
 - Documents (144)
 - Entities (7)
 - Event (1)
 - Organization (0)
 - Person (3)
 - Publication (0)
 - Artifact (0)
 - Place (5)
 - Tag (3)

Note 5: Green Cadres Notes

Entities (12) Status (Open) Assigned Users

Green Cadres Notes

Note Description [Read Only] --- click here for Edit mode

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 were 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?

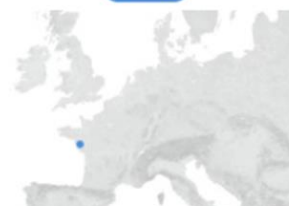
Most Common Person **FRAPET, Guillaume**

Most Common Place **Nantes** 128 docs

Most Recent **Date: 1711/1/29** 1711-1-29

Oldest **Date: 1669/6/5** 1669-6-5

Most Common Place **Nantes** 128 docs



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

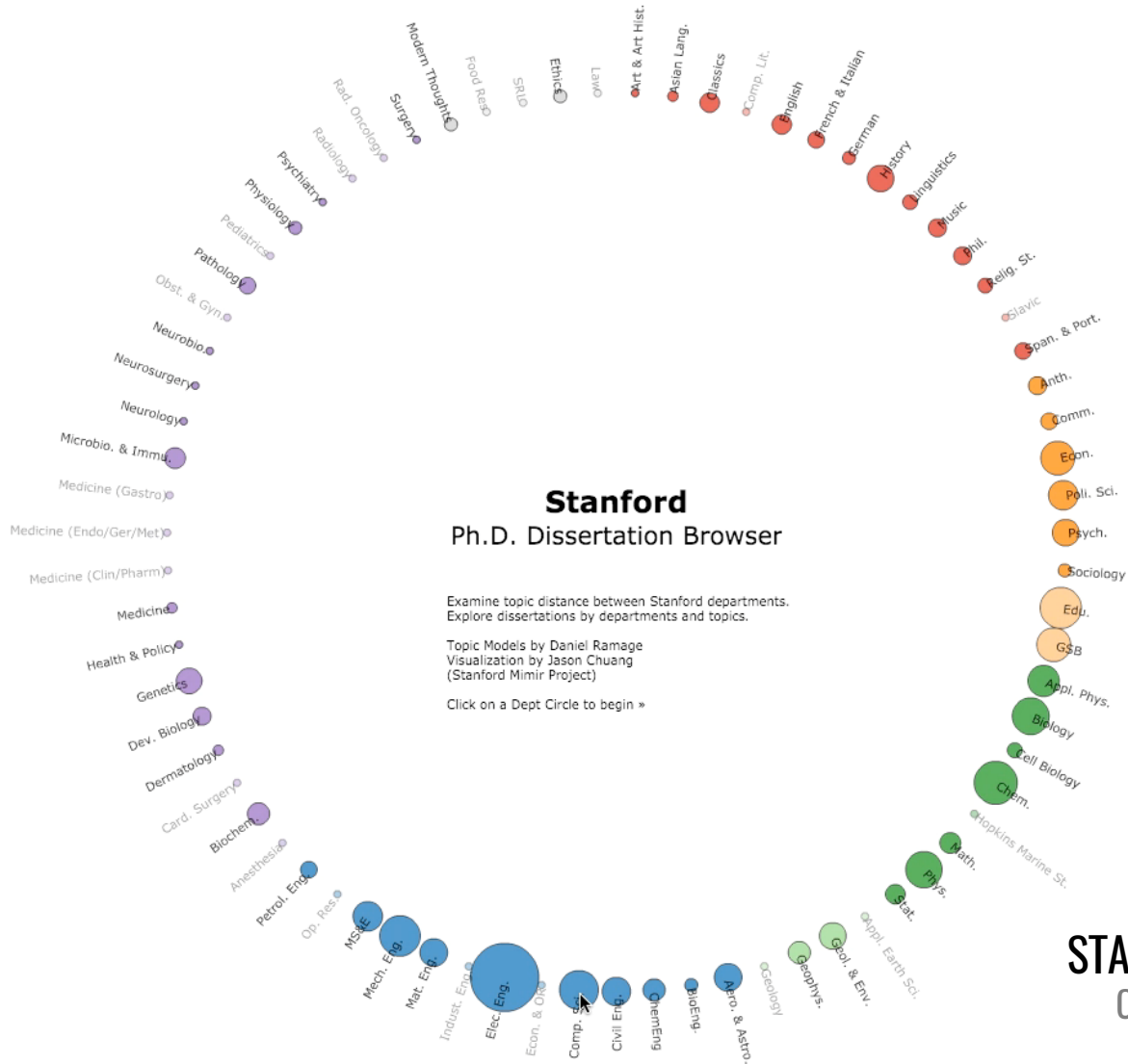
- ASSUME DOCUMENTS ARE A MIXTURE OF TOPICS
 - TOPICS ARE (ROUGHLY) A SET OF CO-OCCURRING TERMS
 - LATENT SEMANTIC ANALYSIS (LSA): REDUCE TERM MATRIX
-
- MANY, MANY APPROACHES EXIST

Stanford Ph.D. Dissertation Browser

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

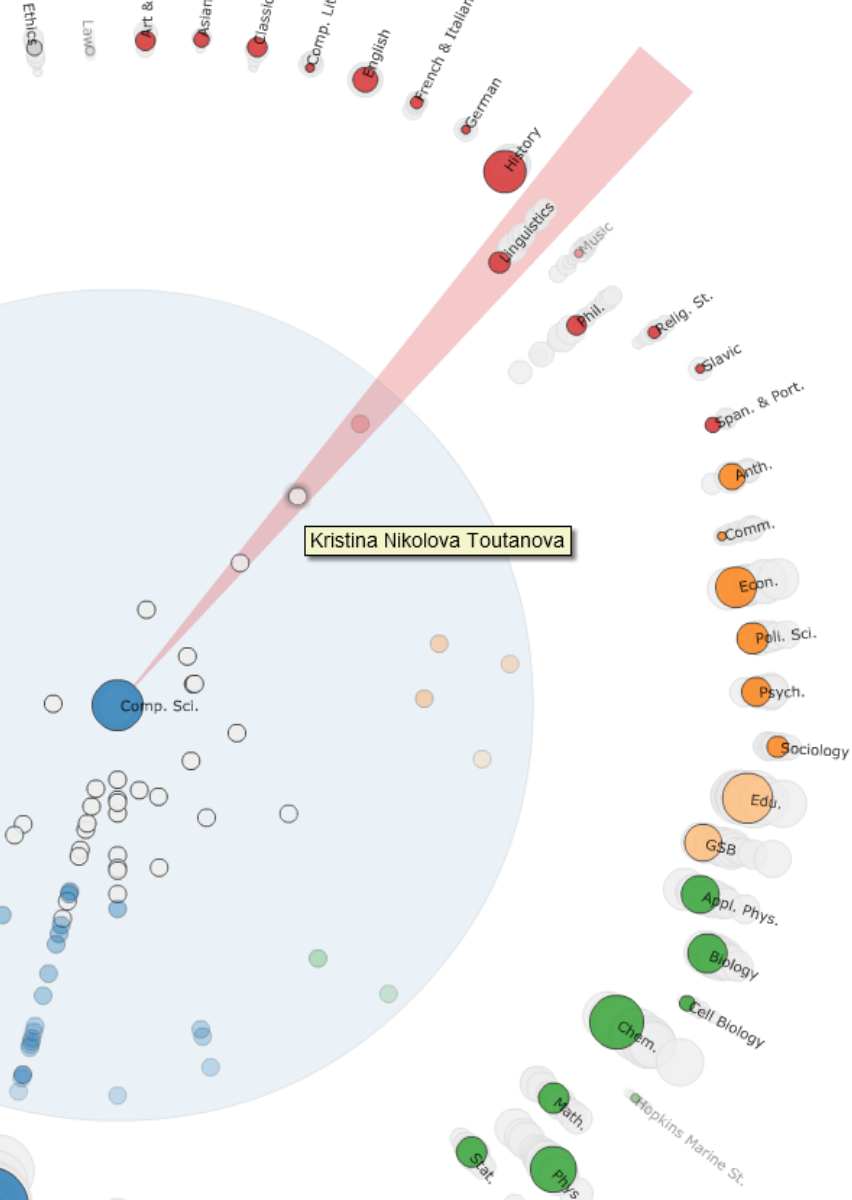
Topic Models by Daniel Ramage
Visualization by Jason Chuang
(Stanford Mimir Project)

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.

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/>

The screenshot shows a web browser window with the URL `textvis.lnu.se`. The page title is "Text Visualization Browser" and the subtitle is "A Visual Survey of Text Visualization Techniques". It is provided by the ISOVIS group. The interface includes a search bar, a time filter set to 2016, and a grid of 272 visualization techniques. A tooltip highlights the "Visual Plagiarism Analysis Tool (2015)".

Text Visualization Browser
A Visual Survey of Text Visualization Techniques
Provided by ISOVIS group

Techniques displayed: **272**

Search:

Time filter: 1976 2016

Analytic Tasks

- Sum
- Alert
- Heart
- Like
- Bell
- Share
- Print
- Refresh
- Edit

Visual Plagiarism Analysis Tool (2015)

Display a menu

QUESTIONS?

EXAM

- 2h, Dec 8th
- bring a pencil
- questions from lectures (at least 1 per lecture)
- some creativity questions
- some questions about assessing visualizations
- every student gets individual exam sheet



+1/1/60+

Introduction to Human-Computer Interaction

Exam on 23/03/2016

- Time period: 8:00 – 11:00
- Duration of the exam: 180 min
- Number of pages: 8
- Materials allowed: Pencils, erasers

Please write your answers directly on the exam paper.

<input type="checkbox"/>	0	<input type="checkbox"/>	0	<input type="checkbox"/>	0	<input type="checkbox"/>	0	<input type="checkbox"/>	0	<input type="checkbox"/>	0	<input type="checkbox"/>	0	<input type="checkbox"/>	0	<input type="checkbox"/>	0	<input type="checkbox"/>	0
<input type="checkbox"/>	1	<input type="checkbox"/>	1	<input type="checkbox"/>	1	<input type="checkbox"/>	1	<input type="checkbox"/>	1	<input type="checkbox"/>	1	<input type="checkbox"/>	1	<input type="checkbox"/>	1	<input type="checkbox"/>	1	<input type="checkbox"/>	1
<input type="checkbox"/>	2	<input type="checkbox"/>	2	<input type="checkbox"/>	2	<input type="checkbox"/>	2	<input type="checkbox"/>	2	<input type="checkbox"/>	2	<input type="checkbox"/>	2	<input type="checkbox"/>	2	<input type="checkbox"/>	2	<input type="checkbox"/>	2
<input type="checkbox"/>	3	<input type="checkbox"/>	3	<input type="checkbox"/>	3	<input type="checkbox"/>	3	<input type="checkbox"/>	3	<input type="checkbox"/>	3	<input type="checkbox"/>	3	<input type="checkbox"/>	3	<input type="checkbox"/>	3	<input type="checkbox"/>	3
<input type="checkbox"/>	4	<input type="checkbox"/>	4	<input type="checkbox"/>	4	<input type="checkbox"/>	4	<input type="checkbox"/>	4	<input type="checkbox"/>	4	<input type="checkbox"/>	4	<input type="checkbox"/>	4	<input type="checkbox"/>	4	<input type="checkbox"/>	4
<input type="checkbox"/>	5	<input type="checkbox"/>	5	<input type="checkbox"/>	5	<input type="checkbox"/>	5	<input type="checkbox"/>	5	<input type="checkbox"/>	5	<input type="checkbox"/>	5	<input type="checkbox"/>	5	<input type="checkbox"/>	5	<input type="checkbox"/>	5
<input type="checkbox"/>	6	<input type="checkbox"/>	6	<input type="checkbox"/>	6	<input type="checkbox"/>	6	<input type="checkbox"/>	6	<input type="checkbox"/>	6	<input type="checkbox"/>	6	<input type="checkbox"/>	6	<input type="checkbox"/>	6	<input type="checkbox"/>	6
<input type="checkbox"/>	7	<input type="checkbox"/>	7	<input type="checkbox"/>	7	<input type="checkbox"/>	7	<input type="checkbox"/>	7	<input type="checkbox"/>	7	<input type="checkbox"/>	7	<input type="checkbox"/>	7	<input type="checkbox"/>	7	<input type="checkbox"/>	7
<input type="checkbox"/>	8	<input type="checkbox"/>	8	<input type="checkbox"/>	8	<input type="checkbox"/>	8	<input type="checkbox"/>	8	<input type="checkbox"/>	8	<input type="checkbox"/>	8	<input type="checkbox"/>	8	<input type="checkbox"/>	8	<input type="checkbox"/>	8
<input type="checkbox"/>	9	<input type="checkbox"/>	9	<input type="checkbox"/>	9	<input type="checkbox"/>	9	<input type="checkbox"/>	9	<input type="checkbox"/>	9	<input type="checkbox"/>	9	<input type="checkbox"/>	9	<input type="checkbox"/>	9	<input type="checkbox"/>	9

← Encode your student number here, and write the student number again as well as your given name and family name below. If you cannot remember your student number, use the number X you see at the top of the exam sheet in this code +X/Y/Z+.

Student number:
.....
Given name:
.....
Family name:
.....

- The questions with the symbol ♣ can have none, one, or more than one possible correct answers. All other questions have exactly one correct answer.
- Please answer the questions like this: ☒ use a **pencil** (hardness HB), and make clear marks. To correct, clearly erase the wrong mark and put a new one (if needed). If you cannot erase because you did not bring a pencil, make the incorrect box completely black.
- All multiple-choice questions are worth one point. For it to be counted as answered correctly, all correct answers and no incorrect answer have to be selected.
- Do not fold the answer sheet(s), do not write on the back.

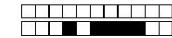
Question 1 Student did **NOT** bring a pencil. Do **NOT** fill out yourself.

- Student brought a pencil.
- Student did not bring a pencil.

Multiple-Choice Questions:

Question 2 Driving to the supermarket but ending up at work is an example of which type of error

- description error
- a mistake
- capture error
- none of the above
- mode error



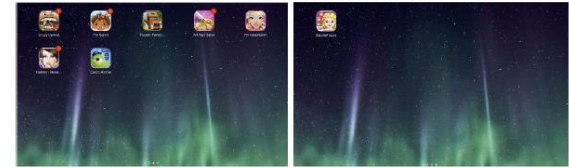
+1/5/56+

Controlled Experiments

You are a designer for a mobile phone company and are trying to decide which method you want people to use for opening apps in future versions of your mobile UI. You are planning to go with either a single page with folders (Interface 1), or multiple scrolling pages and no folders (Interface 2). The choice needs to be made based on which interaction technique allows the user to open an app the fastest. You decide to run a controlled lab experiment to find out.



Interface 1: The main interface shows a single page with folders of icons (left). Clicking on a folder, opens up the folder to show its contents (right).



Interface 2: The main interface shows a page with icons for apps (left). Swiping the page to the side shows a second page with more icons for apps (right).

Question 21 Write an appropriate **hypothesis** for this study (1 point): w c *Reserved*

.....
.....

Question 22 Continuing with the example from the previous question: Write an appropriate **null hypothesis** for this study. w c *Reserved*

.....

EXAM

- best way to mark a box:
- unacceptable way to mark a box:
- if you make an error – erase your answer



- if you forgot your eraser, mark the box like this

ACKNOWLEDGEMENTS

Slides in were inspired, adapted, taken from slides by

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