

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



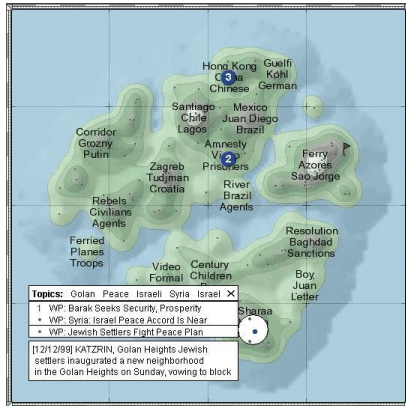
(TODAY)

VISUALIZING TEXT

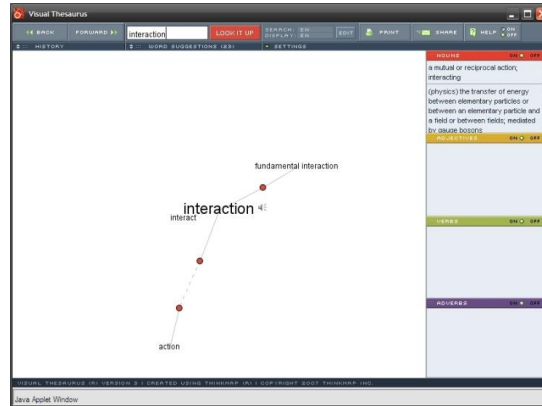
TEXT?

WHY

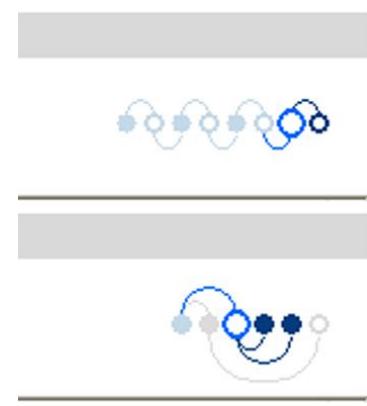
- To assist information retrieval
- To enable linguistic analysis
- To augment analytics on mixed data



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

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)

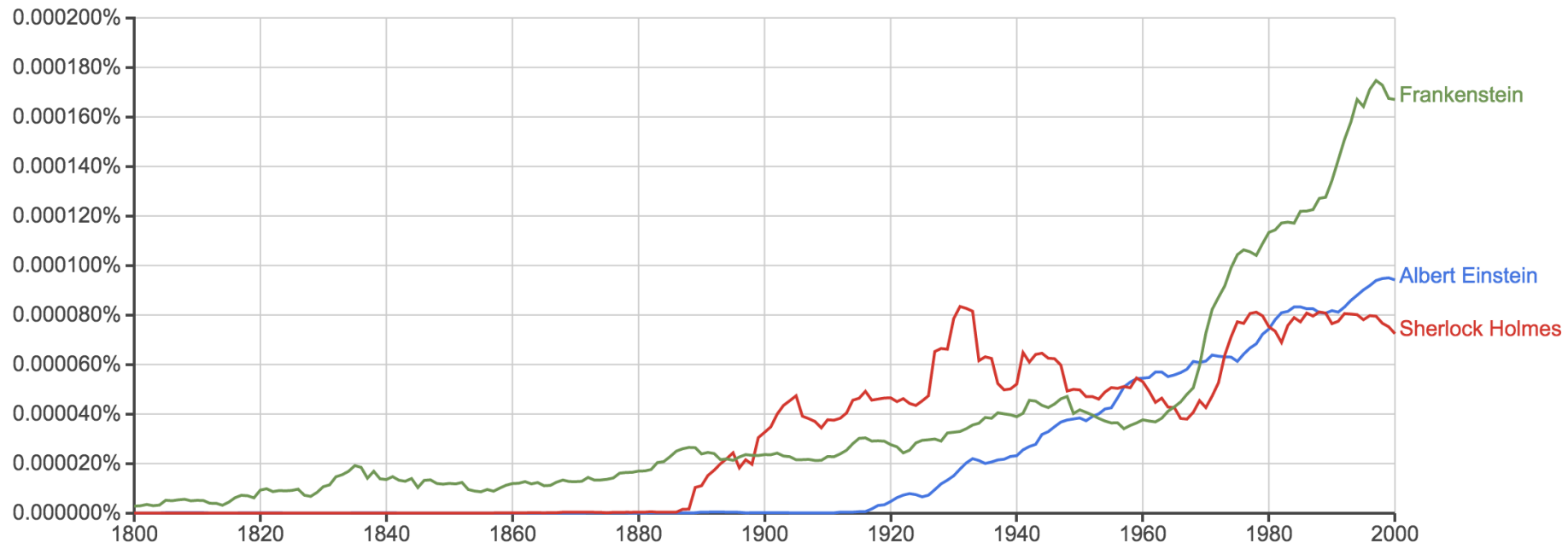
[the quick brown fox jumps], [quick brown fox jumps over], [brown fox jumps over]

...

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**

<http://tagcrowd.com/>

THESIS WESLEY WILLETT

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)

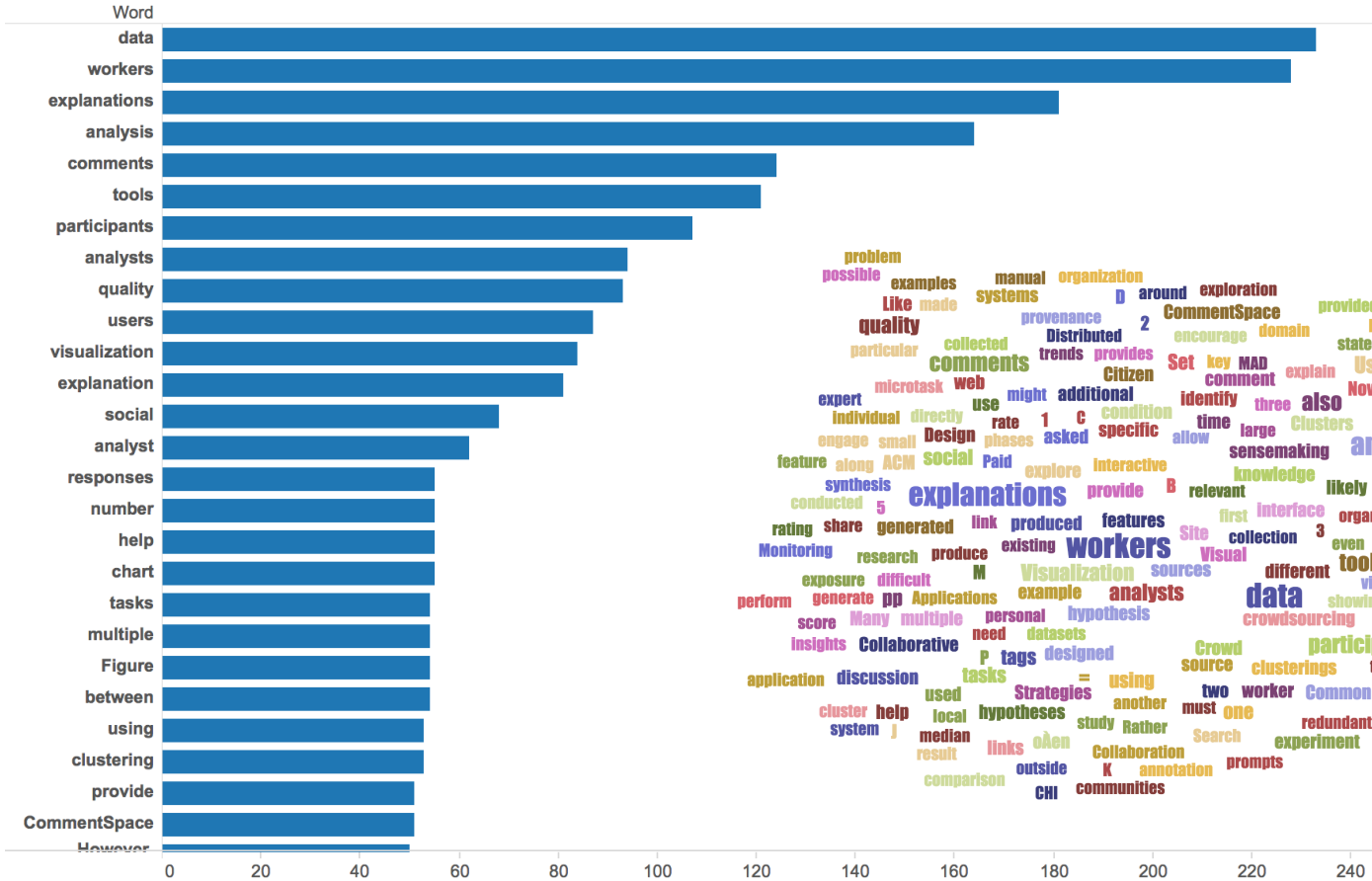
MAY NOT FACILITATE COMPARISON (UNSTABLE LAYOUT)

- ORDER USUALLY MEANINGLESS (USUALLY ALPHABETICAL OR RANDOM)

TERM FREQUENCY MAY NOT BE MEANINGFUL

DOES NOT SHOW THE STRUCTURE OF THE TEXT

WORD COUNTS



WORDCOUNT

WORDCOUNT

◀ PREVIOUS WORD

NEXT WORD ▶

the of and to ain 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'll find it when you're about pen mail they can't

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

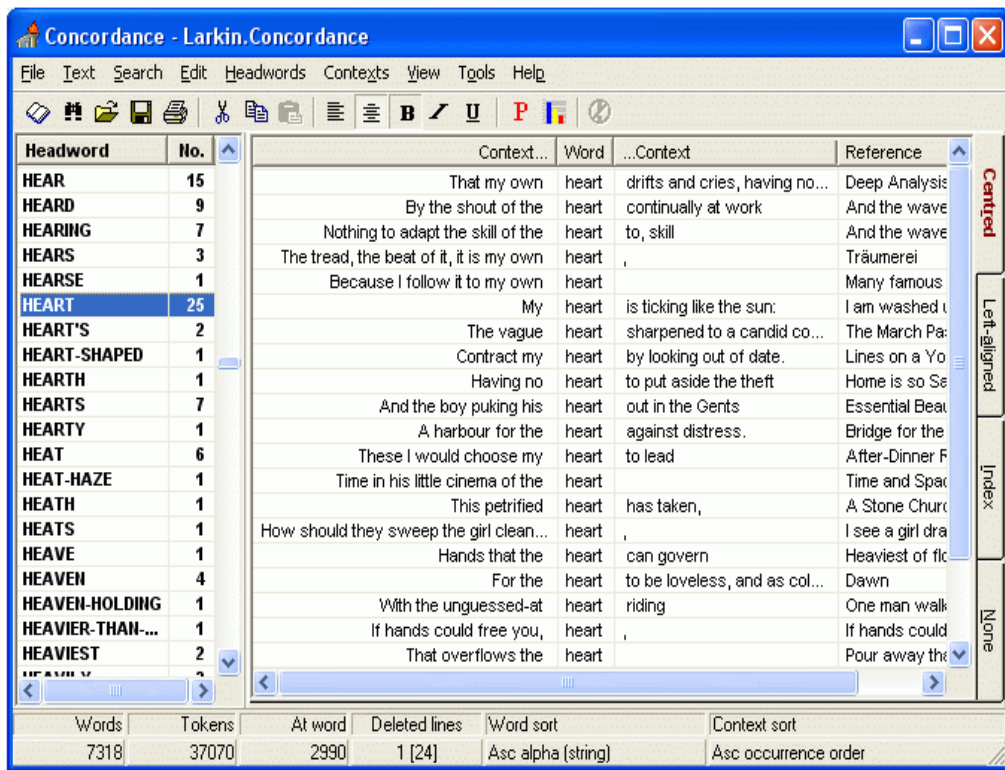


JONATHAN HARRIS

<http://wordcount.org>

CONCORDANCE

WHAT IS THE COMMON LOCAL CONTEXT OF A TERM?



The screenshot shows the 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 various icons, 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), 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 Chur
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 fl
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

thine 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
 with all your heart and with all your soul .

your god

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

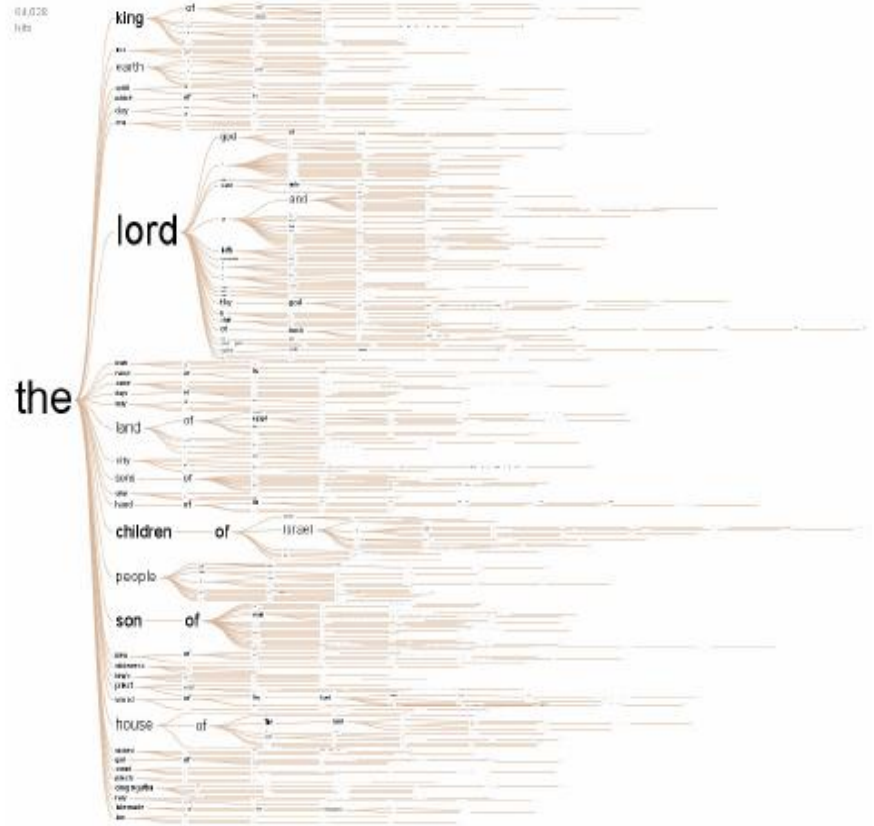
rooms at feasts , and the chief seats in the synagogues , 23 : 7 and greetings in the markets , and to be called of
 seats in the synagogues , and greetings in the markets .

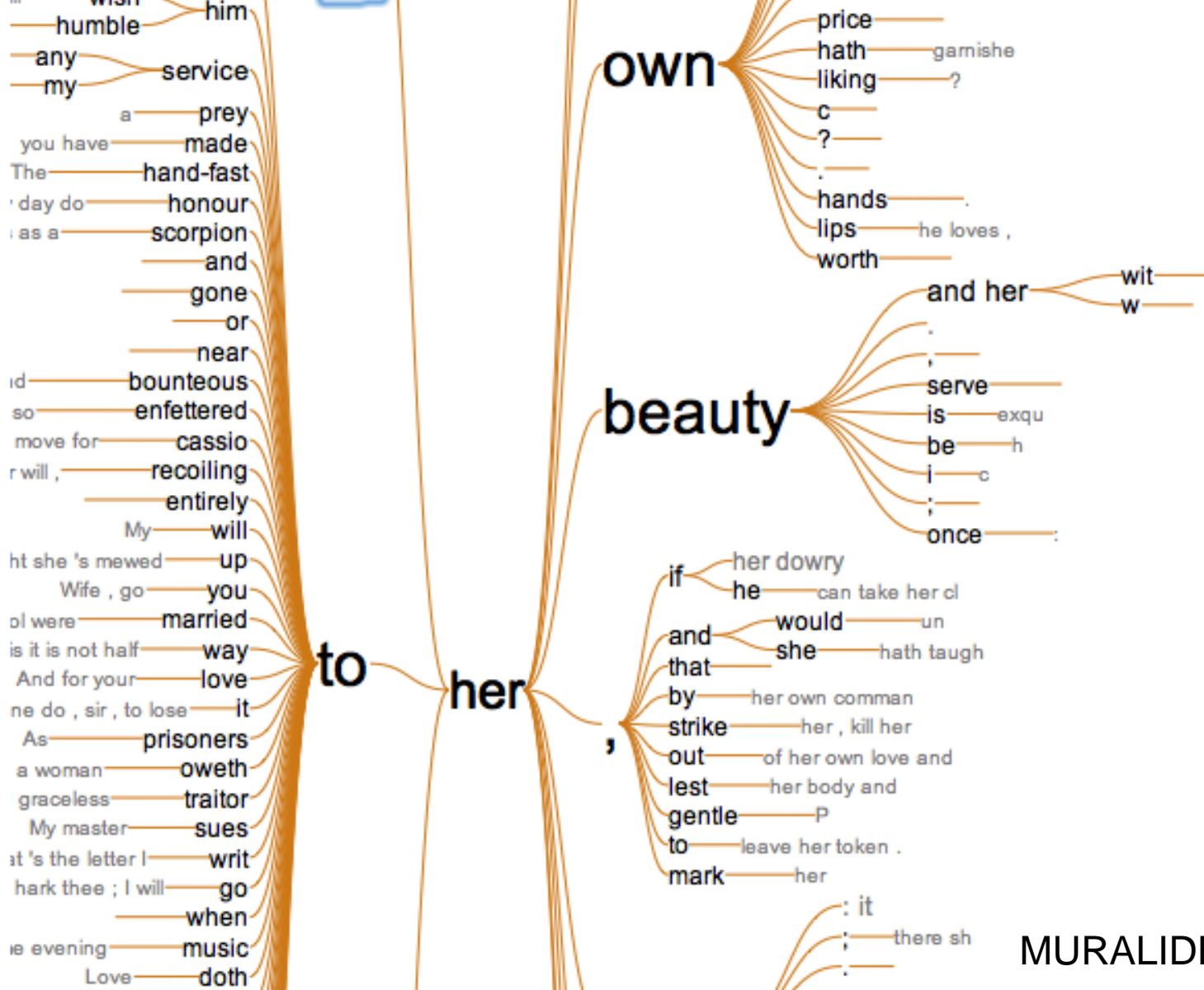
father

; 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

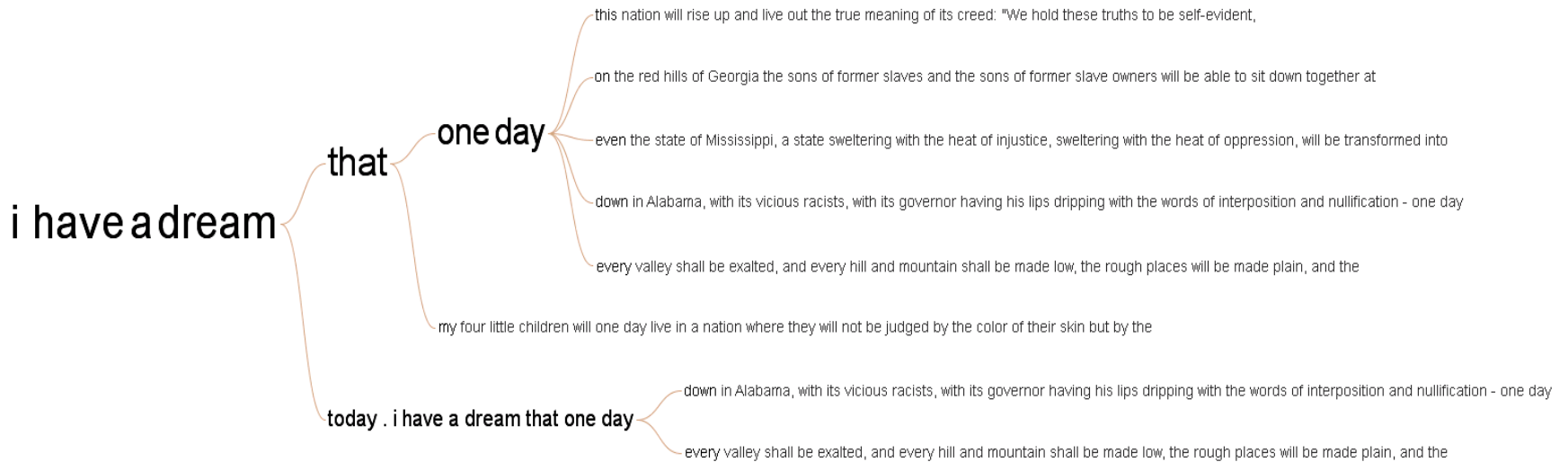
brotherhood .
 world , the love of the father is not in him .
 brethren .
 children of god , when we love god , and keep his commandments .

FILTER INFREQUENT RUNS





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

Showing 73 of 1719 terms

X and Y

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

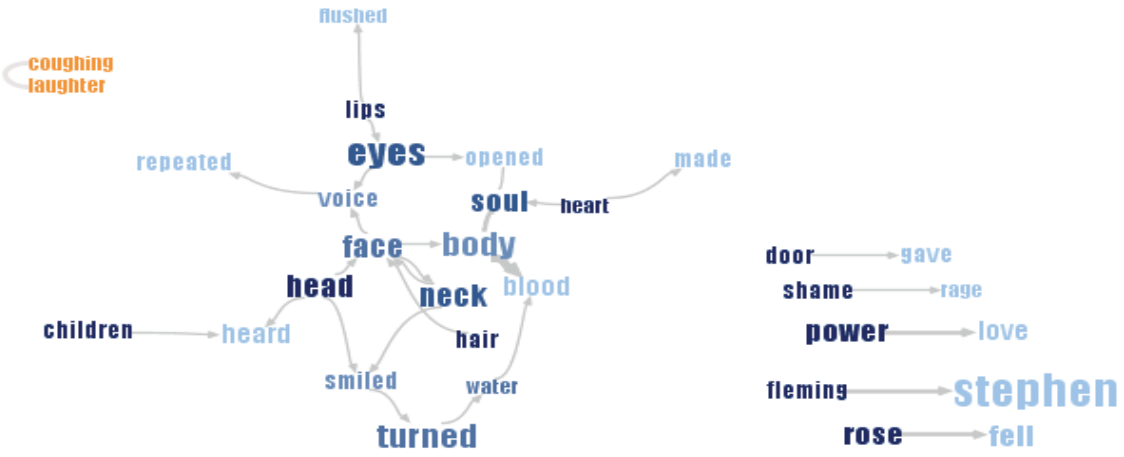
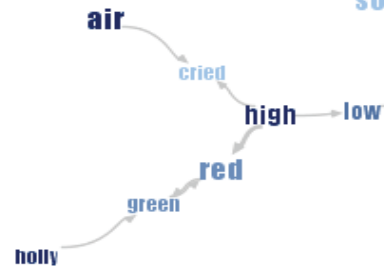
Filters

Show top: 100

Hide common words

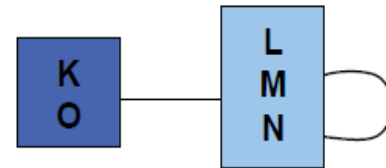
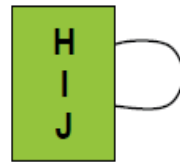
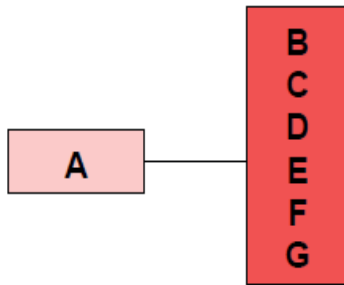
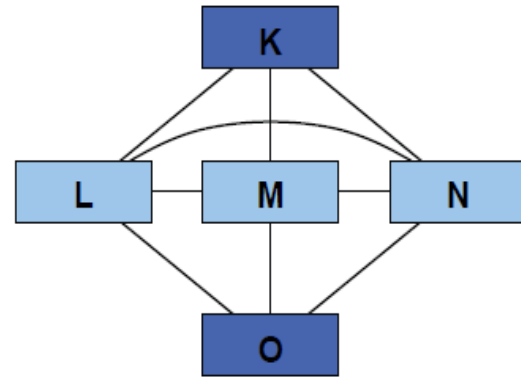
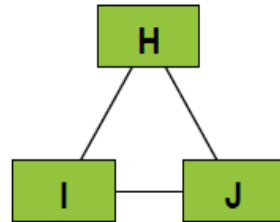
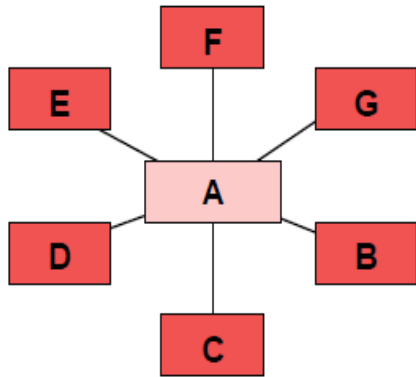
Zoom

In Out Reset



PORTRAIT OF THE ARTIST AS A YOUNG MAN
 JAMES JOYCE

NODE GROUPING



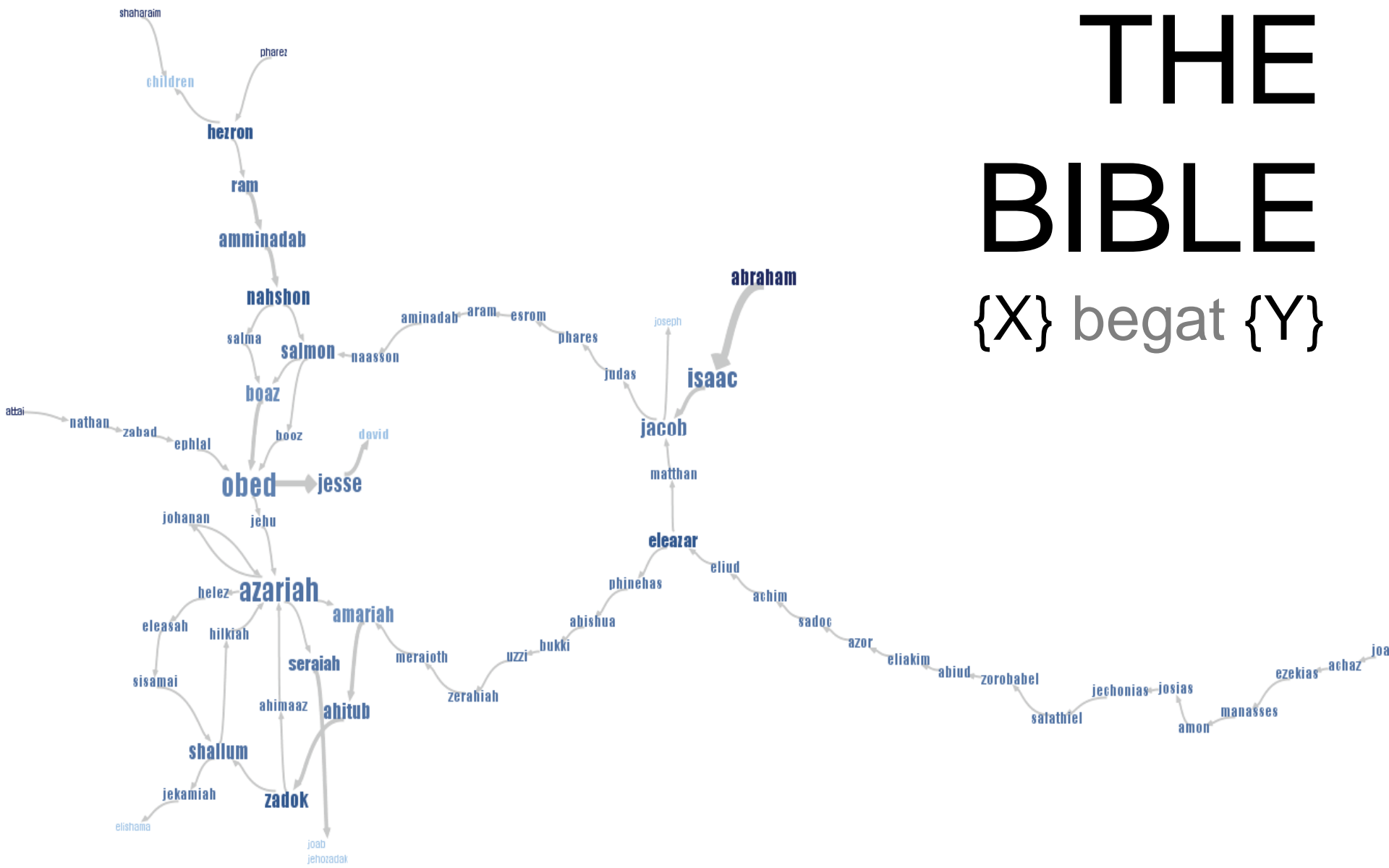
(a)

(b)

(c)

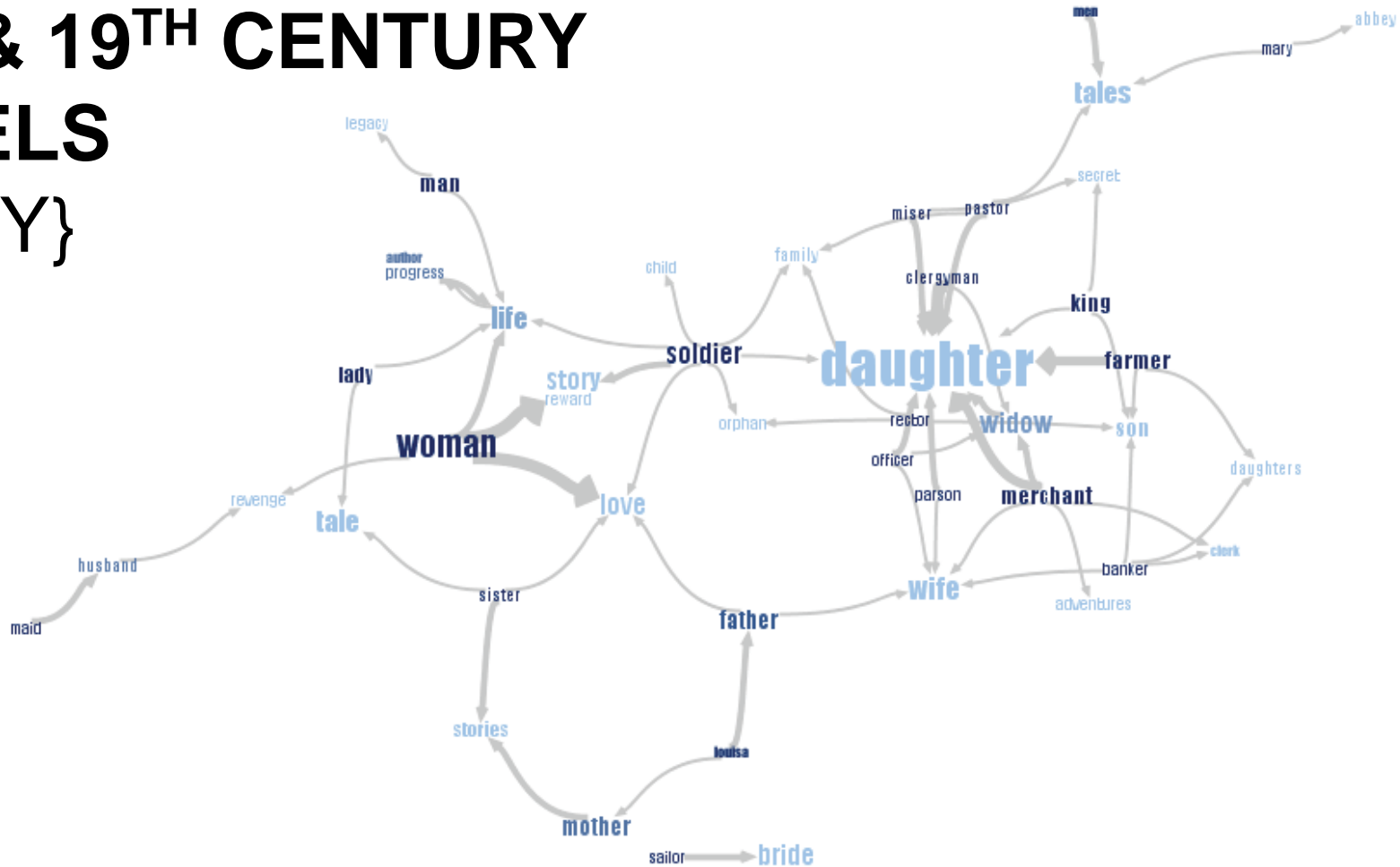
THE BIBLE

{X} begat {Y}



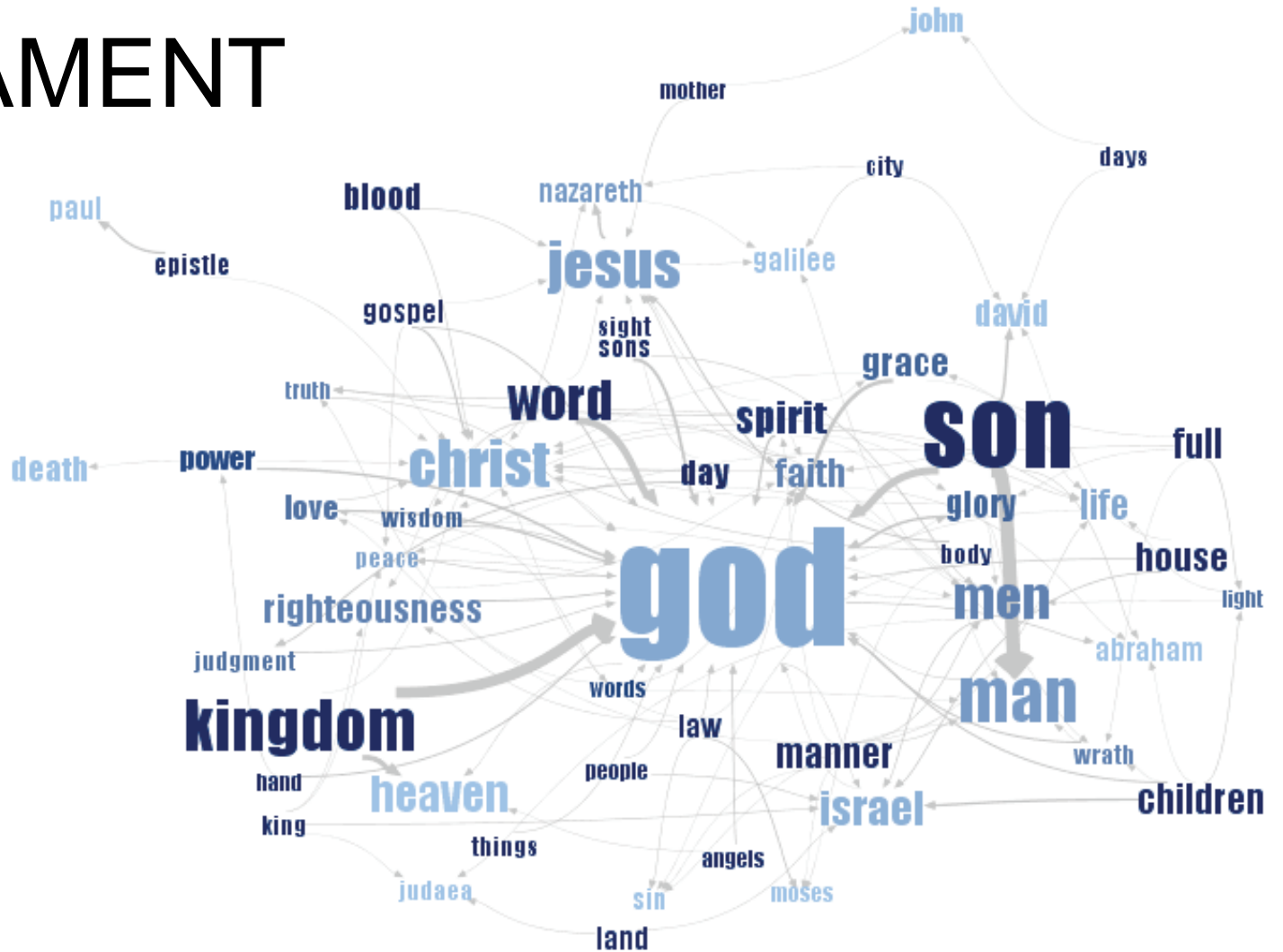
18TH & 19TH CENTURY NOVELS

{X}'s {Y}



NEW TESTAMENT

{X} of {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

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

Highlights from the USC report on entertainment diversity

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

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

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

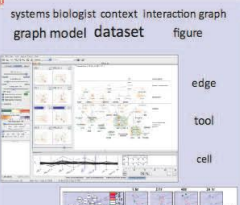
Leafs get set for a busy draft with Matthias trade

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

DOCUMENT CARDS

SMALL MULTIPLES FOR DOCUMENTS

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



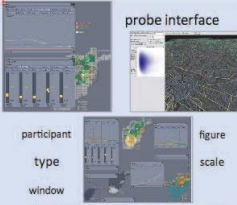
4 systems biologist context interaction graph graph model dataset figure

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



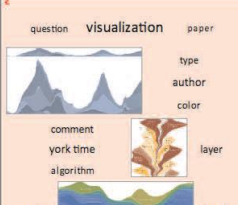
probe interface

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



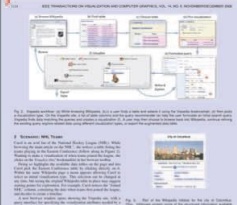
2 question visualization paper

1 2 3 4 5 6 7 8

comment york time algorithm layer 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




3

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



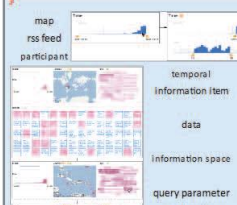
4 edge bundle technique polyline segment

1 2 3 4 5 6 7 8

large graph road map mesh edge pattern transfer function color and opacity enhancement node position control point graph layout result 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



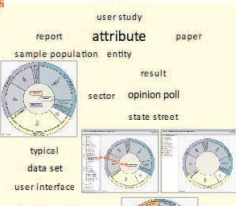
5 map rss feed participant

1 2 3 4 5 6 7 8

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

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

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



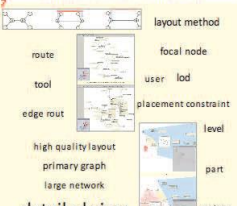
6 user study report attribute paper

1 2 3 4 5 6 7 8

sample population entity result sector opinion poll state street typical data set user interface visualization task design participant poll data system data point

Geoffrey M. Drapes, and Richard F. Riesenfeld

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



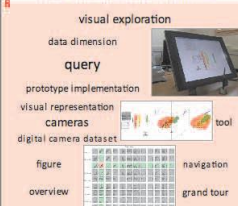
7 layout method focal node

1 2 3 4 5 6 7 8

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

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

Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation



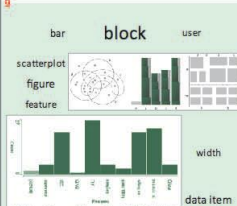
8 visual exploration data dimension query

1 2 3 4 5 6 7 8

prototype implementation visual representation cameras digital camera dataset figure overview range scatterplot matrix user operation method order

Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete

Interactive Visual Analysis of Set-Typed Data



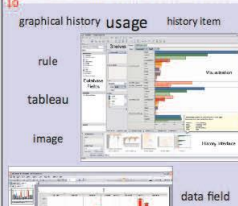
9 bar block user

1 2 3 4 5 6 7 8

scatterplot figure feature width data item dataset data record set-typed data view washing agent histogram

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

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



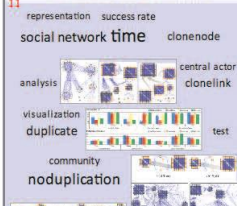
10 graphical history usage history item

1 2 3 4 5 6 7 8

rule tableau image data field display approach event history interface history tool

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

Improving the Readability of Clustered Social Networks using Node Duplication



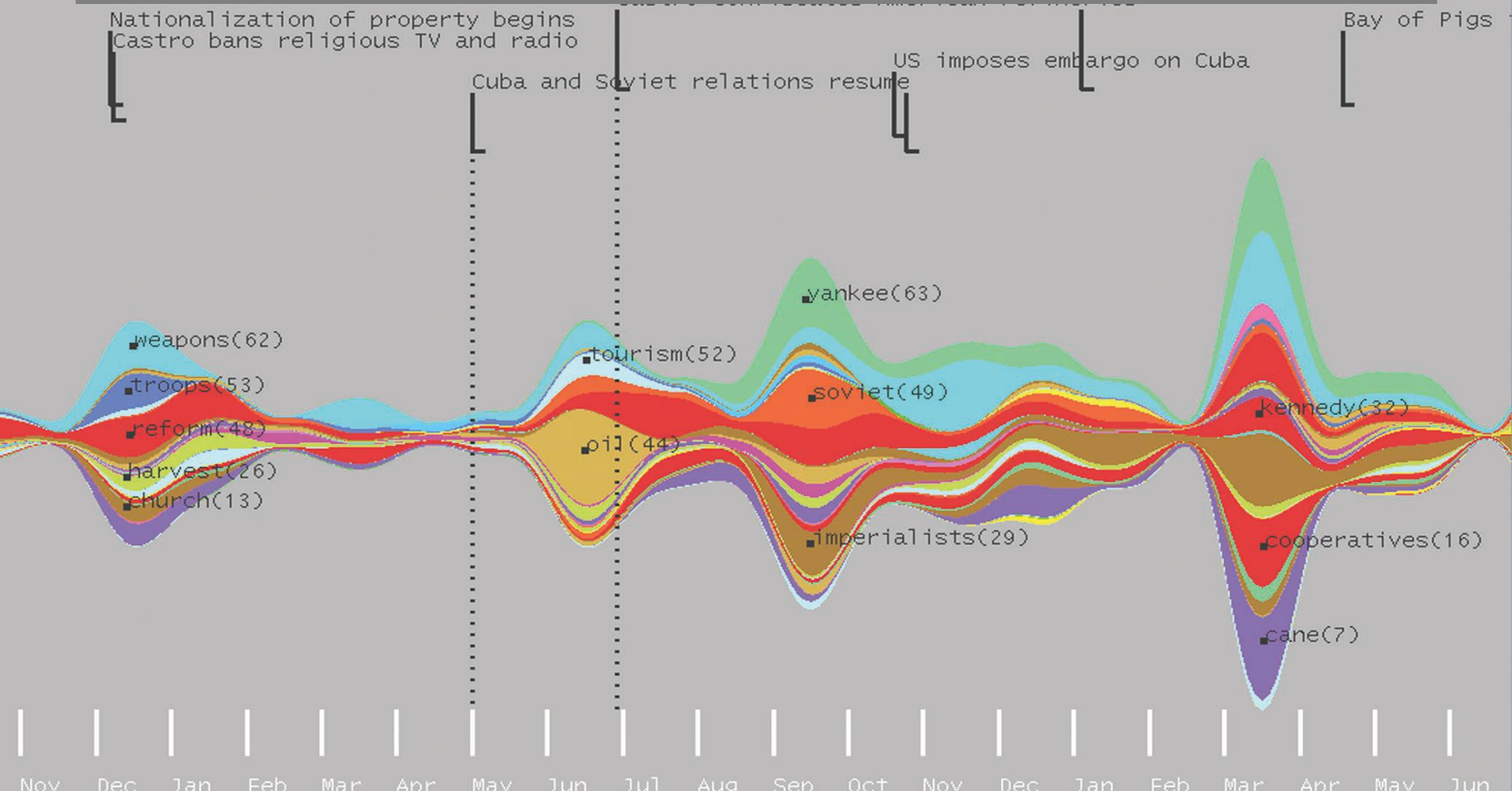
11 representation success rate

1 2 3 4 5 6 7 8

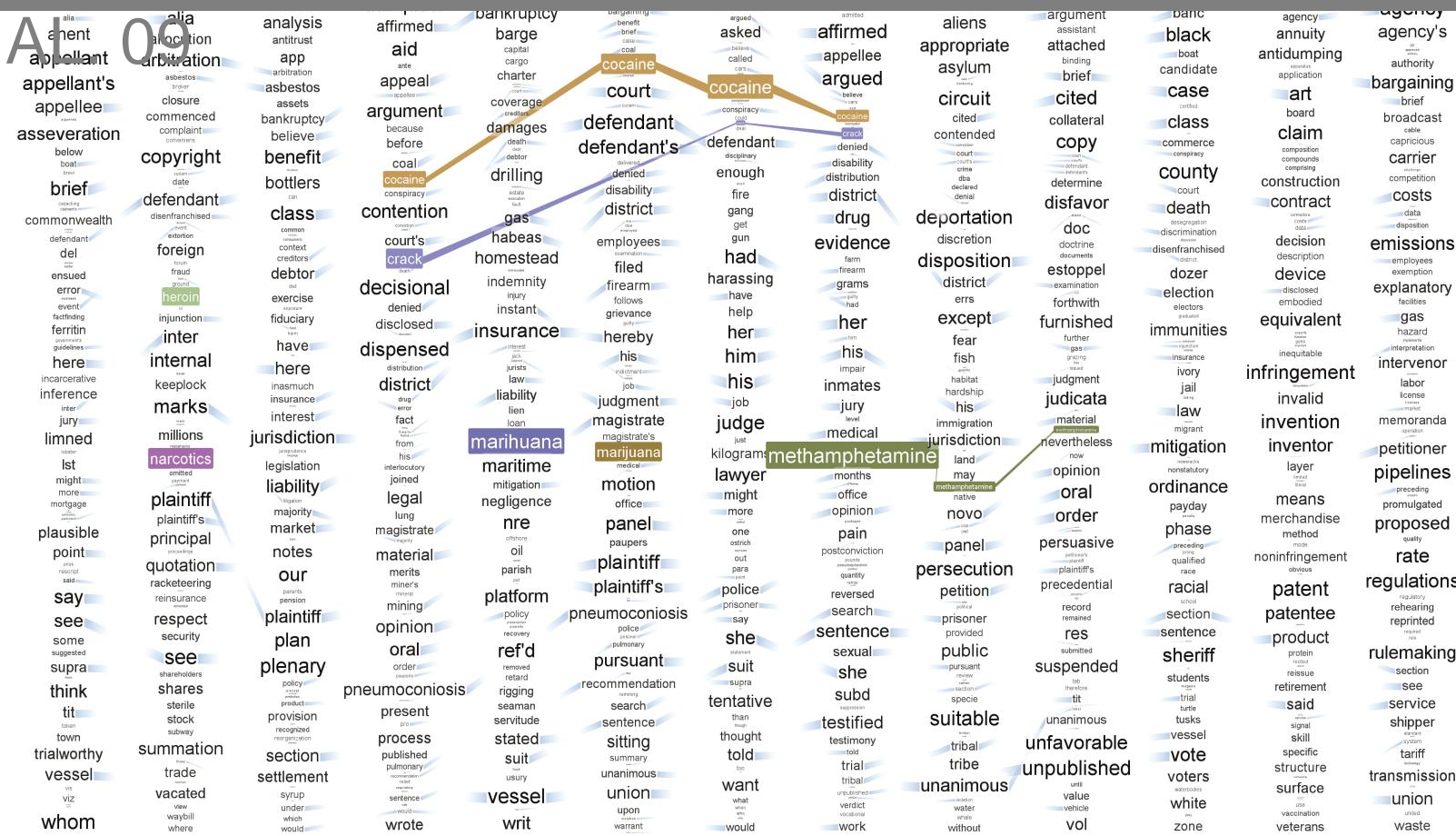
social network time clonemode analysis visualization duplicate community noduplication significant effect cluster duplication link splitlink readability

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

Run Search New Query Quit

Search Limit: 50 100 250 500 1000
Number of Clusters: 3 4 5 8 10

Mode: Tile Bars

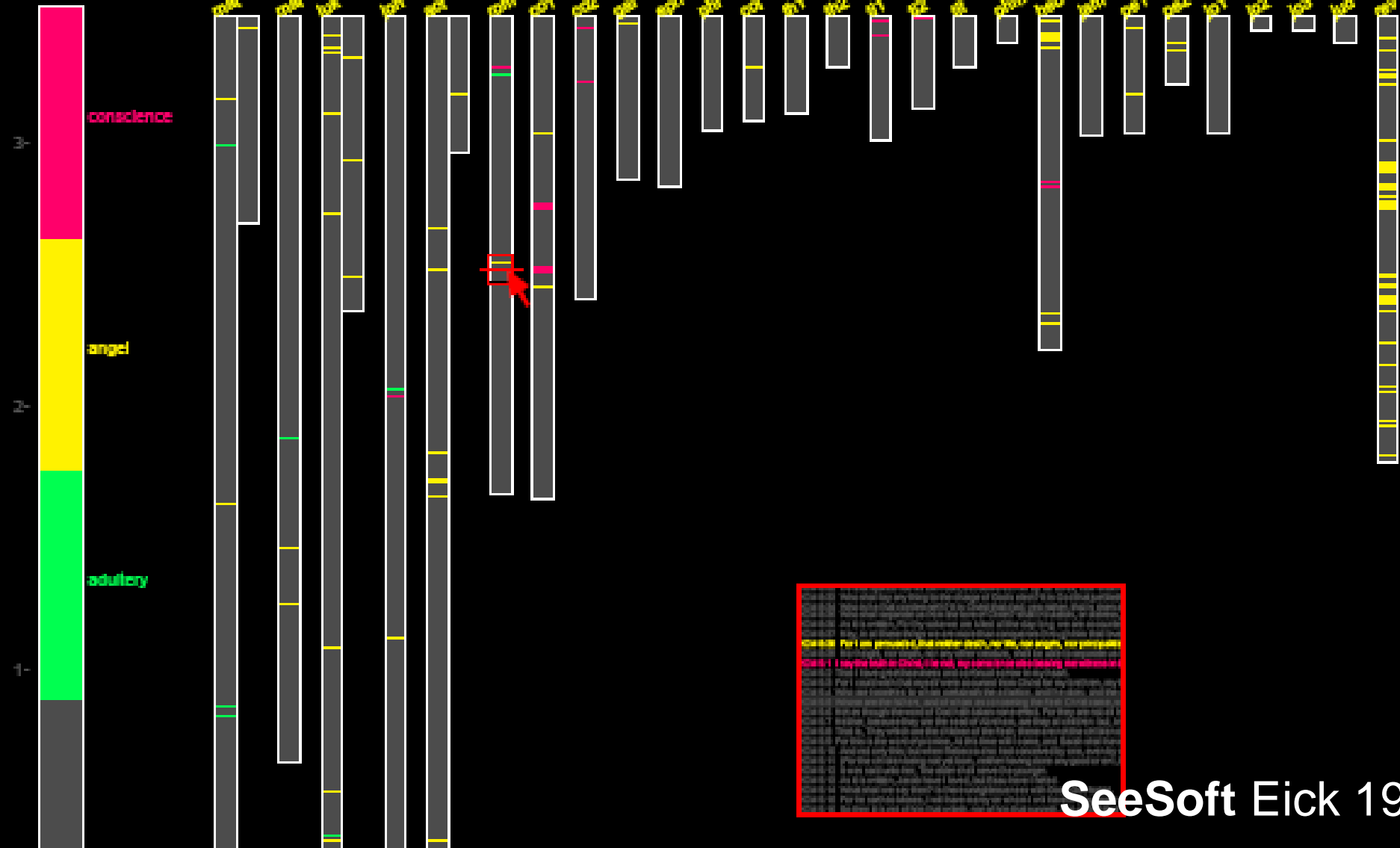
Cluster Titles Backup

The interface displays a list of search results. On the left, a sidebar shows a vertical list of cluster visualizations, each represented by a small grid of colored squares (grey, white, green, black) indicating the distribution of terms across clusters. The main area on the right lists the corresponding search results, including document IDs and titles.

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

#mp/words22058



```

1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040 1041 1042 1043 1044 1045 1046 1047 1048 1049 1050 1051 1052 1053 1054 1055 1056 1057 1058 1059 1060 1061 1062 1063 1064 1065 1066 1067 1068 1069 1070 1071 1072 1073 1074 1075 1076 1077 1078 1079 1080 1081 1082 1083 1084 1085 1086 1087 1088 1089 1090 1091 1092 1093 1094 1095 1096 1097 1098 1099 1100
1101 1102 1103 1104 1105 1106 1107 1108 1109 1110 1111 1112 1113 1114 1115 1116 1117 1118 1119 1120 1121 1122 1123 1124 1125 1126 1127 1128 1129 1130 1131 1132 1133 1134 1135 1136 1137 1138 1139 1140 1141 1142 1143 1144 1145 1146 1147 1148 1149 1150 1151 1152 1153 1154 1155 1156 1157 1158 1159 1160 1161 1162 1163 1164 1165 1166 1167 1168 1169 1170 1171 1172 1173 1174 1175 1176 1177 1178 1179 1180 1181 1182 1183 1184 1185 1186 1187 1188 1189 1190 1191 1192 1193 1194 1195 1196 1197 1198 1199 1200
1201 1202 1203 1204 1205 1206 1207 1208 1209 1210 1211 1212 1213 1214 1215 1216 1217 1218 1219 1220 1221 1222 1223 1224 1225 1226 1227 1228 1229 1230 1231 1232 1233 1234 1235 1236 1237 1238 1239 1240 1241 1242 1243 1244 1245 1246 1247 1248 1249 1250 1251 1252 1253 1254 1255 1256 1257 1258 1259 1260 1261 1262 1263 1264 1265 1266 1267 1268 1269 1270 1271 1272 1273 1274 1275 1276 1277 1278 1279 1280 1281 1282 1283 1284 1285 1286 1287 1288 1289 1290 1291 1292 1293 1294 1295 1296 1297 1298 1299 1300
1301 1302 1303 1304 1305 1306 1307 1308 1309 1310 1311 1312 1313 1314 1315 1316 1317 1318 1319 1320 1321 1322 1323 1324 1325 1326 1327 1328 1329 1330 1331 1332 1333 1334 1335 1336 1337 1338 1339 1340 1341 1342 1343 1344 1345 1346 1347 1348 1349 1350 1351 1352 1353 1354 1355 1356 1357 1358 1359 1360 1361 1362 1363 1364 1365 1366 1367 1368 1369 1370 1371 1372 1373 1374 1375 1376 1377 1378 1379 1380 1381 1382 1383 1384 1385 1386 1387 1388 1389 1390 1391 1392 1393 1394 1395 1396 1397 1398 1399 1400
1401 1402 1403 1404 1405 1406 1407 1408 1409 1410 1411 1412 1413 1414 1415 1416 1417 1418 1419 1420 1421 1422 1423 1424 1425 1426 1427 1428 1429 1430 1431 1432 1433 1434 1435 1436 1437 1438 1439 1440 1441 1442 1443 1444 1445 1446 1447 1448 1449 1450 1451 1452 1453 1454 1455 1456 1457 1458 1459 1460 1461 1462 1463 1464 1465 1466 1467 1468 1469 1470 1471 1472 1473 1474 1475 1476 1477 1478 1479 1480 1481 1482 1483 1484 1485 1486 1487 1488 1489 1490 1491 1492 1493 1494 1495 1496 1497 1498 1499 1500

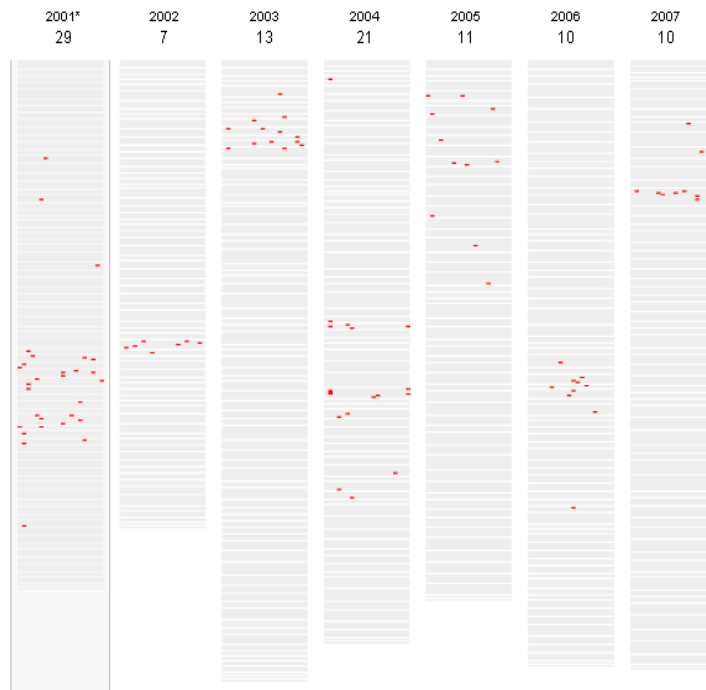
```

SeeSoft Eick 19

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



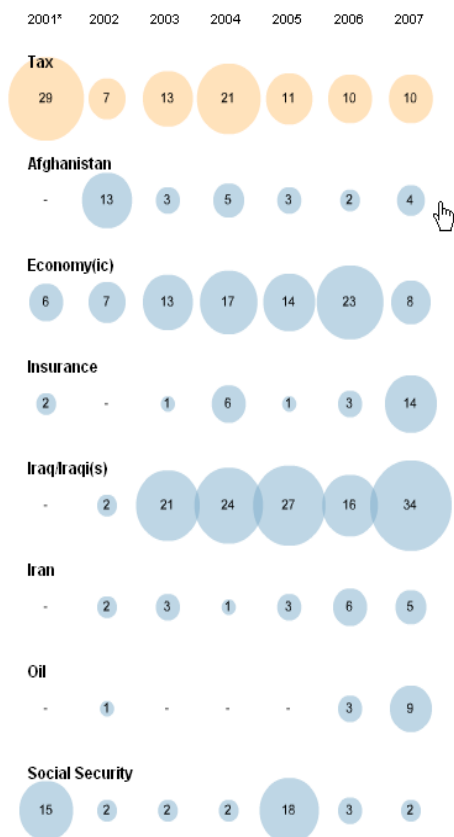
The word in context

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.

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.

Robrdantz, C.

Schroek, T.

Stoffel, A.

Albuquerque, G.

Ankerst, M.

Berchtold, S.

Danon, G.

Deussen, O.

Eisemann, M.

Grise, T.

Haug, L. E.

Heilmann, R.

Hsu, M.

Jenny, M.

Ladisch, J.

Last, M.

Mignock, M.

Muennich, D.

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

mathematics

Document Cluster View

Edit View Bookmarks Export Options

right-viewed documents

Filters

All Filters

Group by Filters

Undo Filters

Hide Unfiltered

Clusters

Text Seed (30)

First Words Unique Words

All Documents

- visual animation,trends: 21
- visualization,users,tablec
- visualization,user,transf
- visualization,use,classif
- visual insight,genes: 19
- visualization,tree,eye: 56
- visualizing,users,space

animation,trends,causality

transforms,quality,studied

night,genes,expression

tables,database,interfaces

classification,geographic,statistics

treemaps,coloring,hierarchicaly

dimensions,coordinates,parallel

network,graph,social

graphs,edge,algorithm

spaces,internet,search

interact,understand,cognition

text,features,topic

video,explorer,stories

collaborative,uses,framework

diverse,environments,toolkit

3d,displays,navigation


history,mining,patterns

analytics,anomalies,detect

querying,series,temporal

state,displayed,explored

JIGSAW



CENDARI
COLLABORATIVE EUROPEAN DIGITAL
ARCHIVE INFRASTRUCTURE

[Home](#) [Browse](#) [About](#) [Issue Report](#) [Survey](#)

[anthi](#)

Resources New Save Import Help

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 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?

Visualizations

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



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

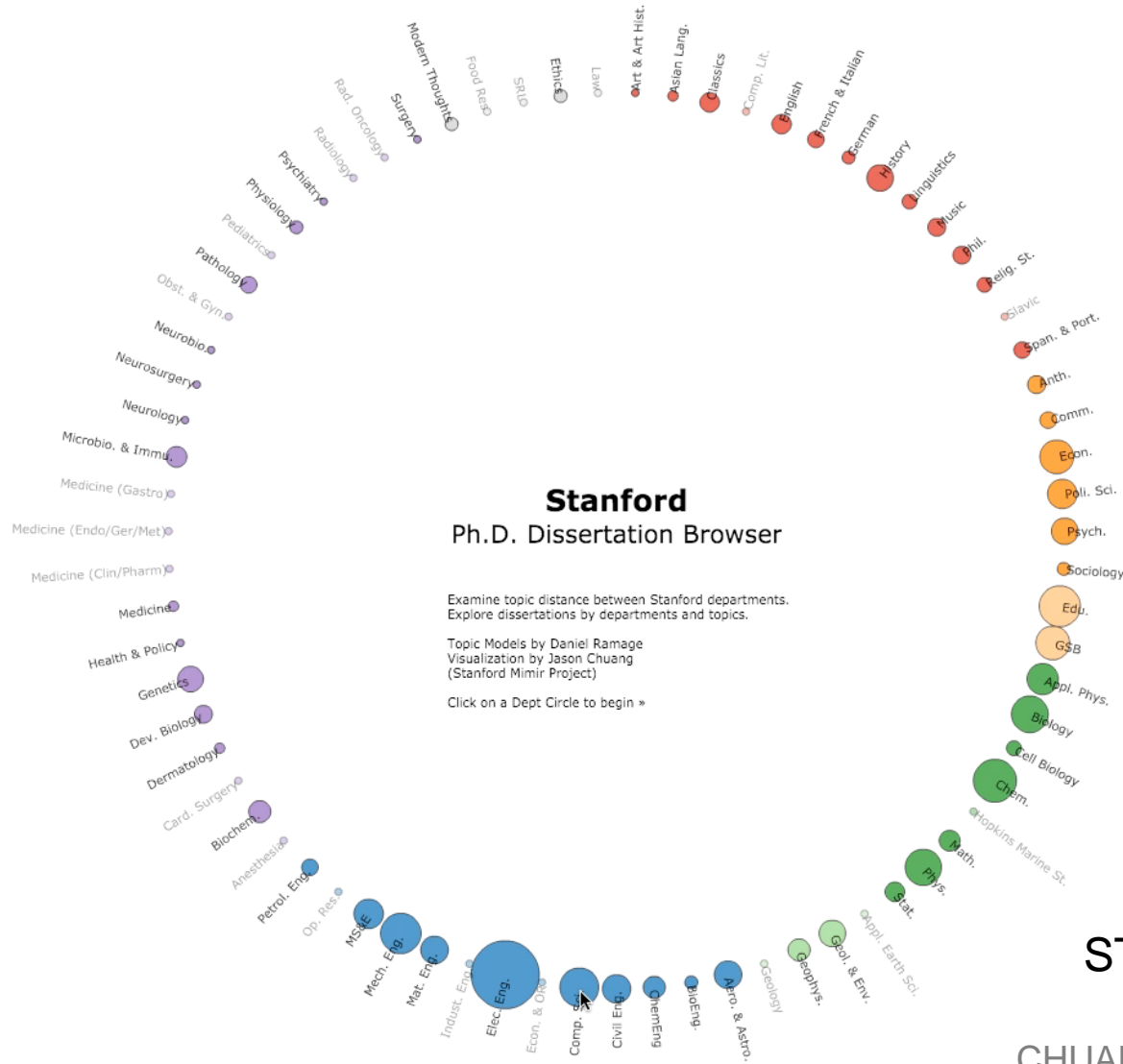
- ASSUME DOCUMENTS ARE A MIXTURE OF TOPICS
 - TOPICS ARE (ROUGHLY) A SET OF CO-OCCURRING TERMS
 - LATENT SEMANTIC ANALYSIS (LSA): REDUCE TERM MATRIX
-
- MANY, MANY APPROACHES EXIST

Stanford Ph.D. Dissertation Browser

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

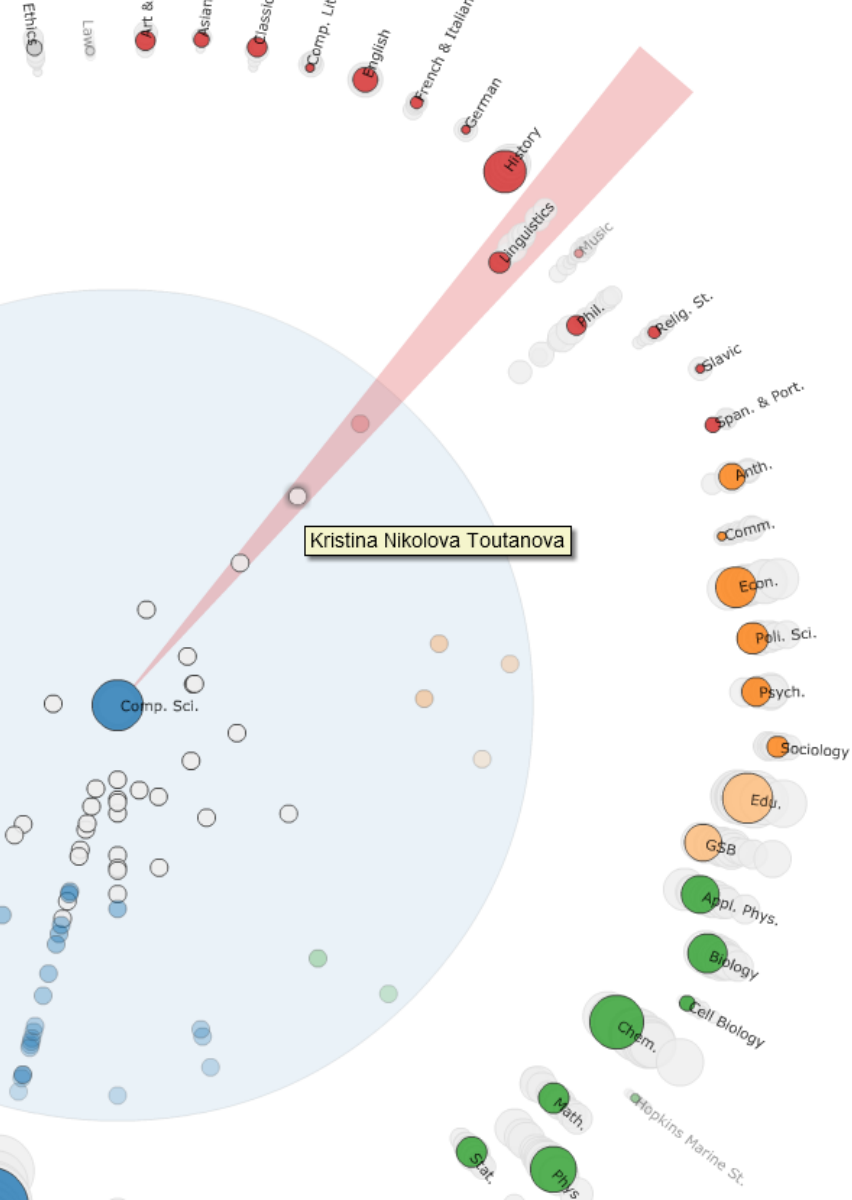
Topic Models by Daniel Ramage
Visualization by Jason Chuang
(Stanford Mimir Project)

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 page features a navigation menu with "About", "Add entry", and "Other surveys".

On the left side, there is a sidebar with the following elements:

- Techniques displayed:** 272
- Search:** A search input field with a clear button (X).
- Time filter:** A range slider from 1976 to 2016.
- Analytic Tasks:** A grid of icons representing different tasks: Sum, Alert, Heart, Like, Bell, Cross, Text, Refresh, and Share.

The main content area displays a grid of 48 text visualization techniques. A tooltip is visible over one of the techniques, labeled "Visual Plagiarism Analysis Tool (2015)". The techniques include various charts, maps, and diagrams, such as word clouds, network graphs, and radar charts.

At the bottom right of the page, there is a button labeled "Display a menu".

QUESTIONS?

ACKNOWLEDGEMENTS

Slides in were inspired, adapted, taken from slides by

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