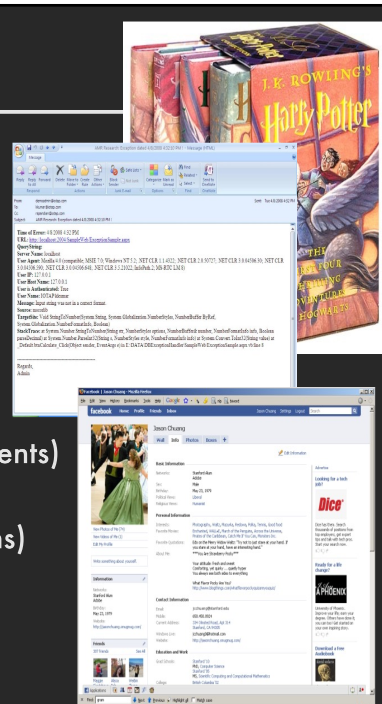
Text Visualization

Maneesh Agrawala

CS 448B: Visualization
Fall 2021

1

Text as data



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)

2

Text Visualization

3

Why visualize text?

4

Why Visualize Text?

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

5

Example: Health Care Reform

Background

- Initiatives by President Clinton
- Overhaul by President Obama

Text data

- News articles
- Speech transcriptions
- Legal documents

What questions might you want to answer?
What visualizations might help?

6

A Concrete Example

September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of you

7

Word/Tag Clouds: Word Count

President Obama's Health Care Speech to Congress

economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

8

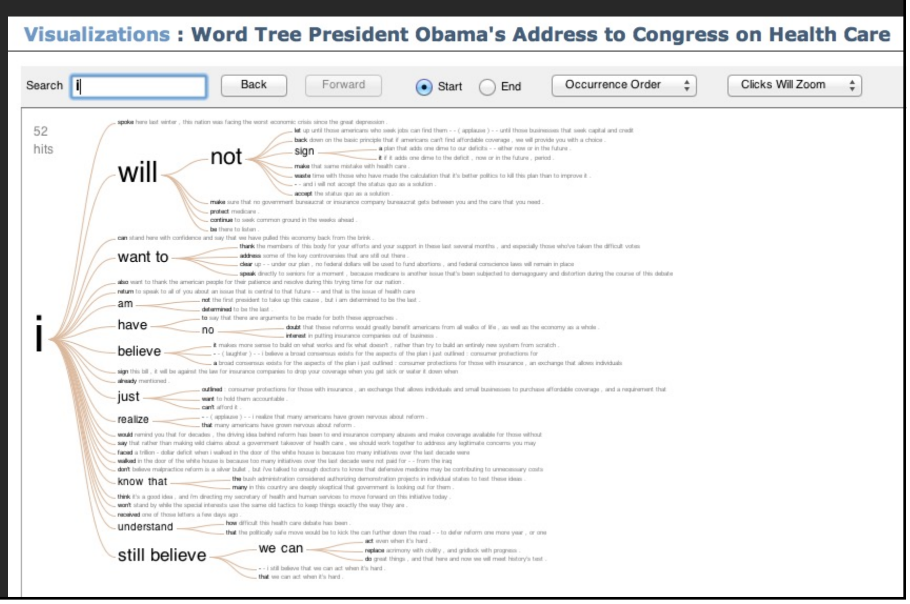
Bill Clinton 1993

Barack Obama 2009

economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

9

WordTree: Word Sequences



10

11

Gulf of Evaluation

Many (most?) text visualizations do not represent text directly. They represent the output of a **language model** (word counts, word sequences, etc.)

Can you interpret the visualization?

How well does it convey the properties of the model?

Do you trust the model?

How does the model enable us to reason about the text?

12

Topics

Text as Data

Visualizing Document Content

Visualizing Conversation

Document Collections

15

Text as Data

16

Words as nominal data?

High dimensional (10,000+)

More than equality tests

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

Words have meanings and relations

17

Text Processing Pipeline

Tokenization

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

18

Text Processing Pipeline

Tokenization

Segment text into terms.

Remove stop words? *a, an, the, of, to, be*

Numbers and symbols? *#cardinal, @Stanford, OMG!!!!!!!!*

Entities? *Palo Alto, O'Connor, U.S.A.*

Stemming

Group together different forms of a word.

Porter stemmer? *visualization(s), visualize(s), visually* -> *visual*

Lemmatization? *goes, went, gone* -> *go*

19

Text Processing Pipeline

Tokenization

Segment text into terms.

Remove stop words? *a, an, the, of, to, be*

Numbers and symbols? *#cardinal, @Stanford, OMG!!!!!!!!*

Entities? *Palo Alto, O'Connor, U.S.A.*

Stemming

Group together different forms of a word.

Porter stemmer? *visualization(s), visualize(s), visually* -> *visual*

Lemmatization? *goes, went, gone* -> *go*

Ordered list of terms

20

The Bag of Words Model

Ignore ordering relationships within the text

A document ≈ vector of term weights

Each term corresponds to a dimension (10,000+)

Each value represents the relevance

- For example, simple term counts

Aggregate into a document x term matrix

Document vector space model

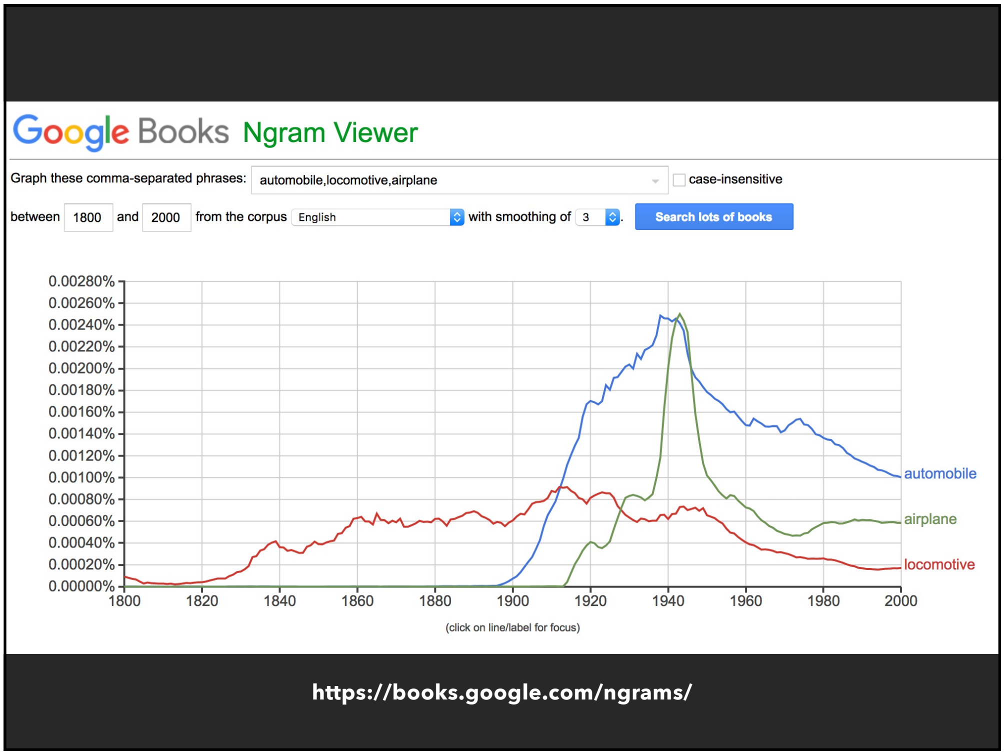
21

Document x Term matrix

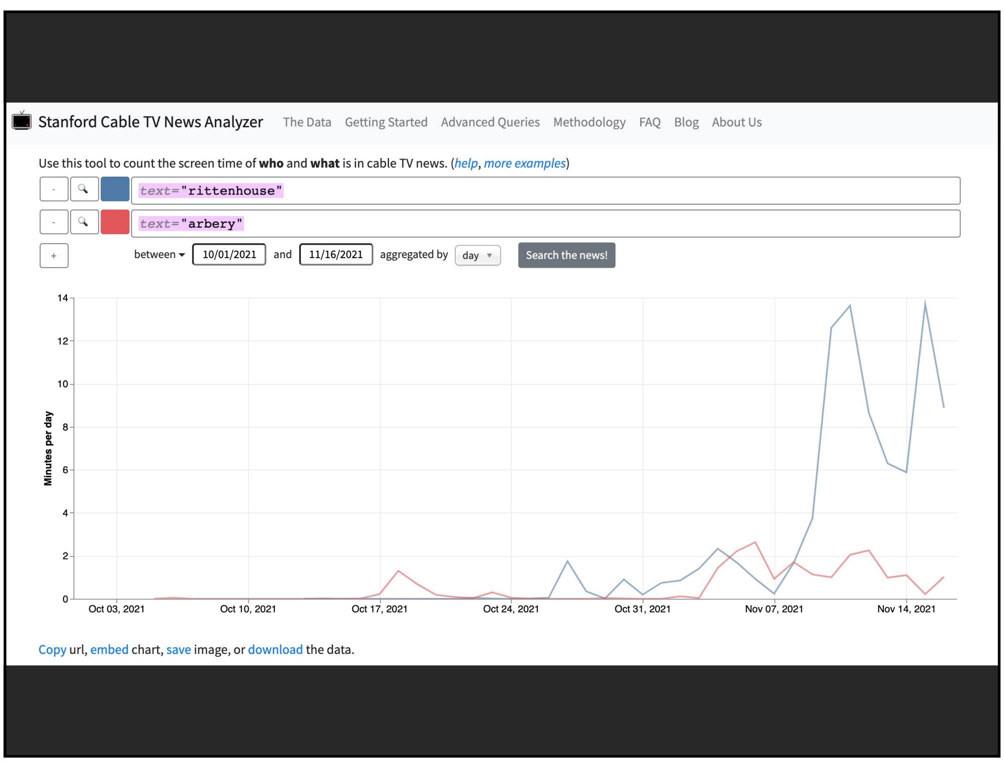
Each document is a vector of term weights
Simplest weighting is to just count occurrences

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

22


Google Books Ngram Viewer
Graph these comma-separated phrases: automobile,locomotive,airplane
case-insensitive
between 1800 and 2000 from the corpus English with smoothing of 3.
Search lots of books
automobile
airplane
locomotive
(click on line/label for focus)
https://books.google.com/ngrams/

24


Stanford Cable TV News Analyzer
The Data
Getting Started
Advanced Queries
Methodology
FAQ
Blog
About Us
Use this tool to count the screen time of who and what is in cable TV news. (help, more examples)
text="rittenhouse"
text="arbery"
between 10/01/2021 and 11/16/2021 aggregated by day
Search the news!
Minutes per day
Oct 03, 2021
Oct 10, 2021
Oct 17, 2021
Oct 24, 2021
Oct 31, 2021
Nov 07, 2021
Nov 14, 2021
Copy url, embed chart, save image, or download the data.

26

27

Word/Tag Clouds

Strengths

Can help with gisting and initial query formation

Weaknesses

Sub-optimal visual encoding (size not pos. encodes freq.)
Inaccurate size encoding (long words are bigger)
May not facilitate comparison (unstable layout)
Term frequency may not be meaningful
Does not show the structure of the text

28

Announcements

29

Final project

Data analysis/explainer or conduct research

- **Data analysis**: Analyze dataset in depth & make a visual explainer
- **Research**: Pose problem, Implement creative solution

Deliverables

- **Data analysis/explainer**: Article with multiple different interactive visualizations
- **Research**: Implementation of solution and web-based demo if possible
- **Short video (2 min)** demoing and explaining the project

Schedule

- Project proposal: Wed 11/3
- Design Review and Feedback: 10th week of quarter
- Final code and video: Fri 12/10 11:59pm

Grading

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

30

Feedback (Week 10)

Signup for a ~10 min slot

https://docs.google.com/spreadsheets/d/1U-Q7DVvWTmTt_nYumJlqSDgySvAd2Iq0EuUXGVHo4cE/edit?usp=sharing

Plan to give a 5 min presentation (mostly demo) of work so far. We will give oral feedback.

31

Given a text, what are the best descriptive words?

32

Keyword Weighting

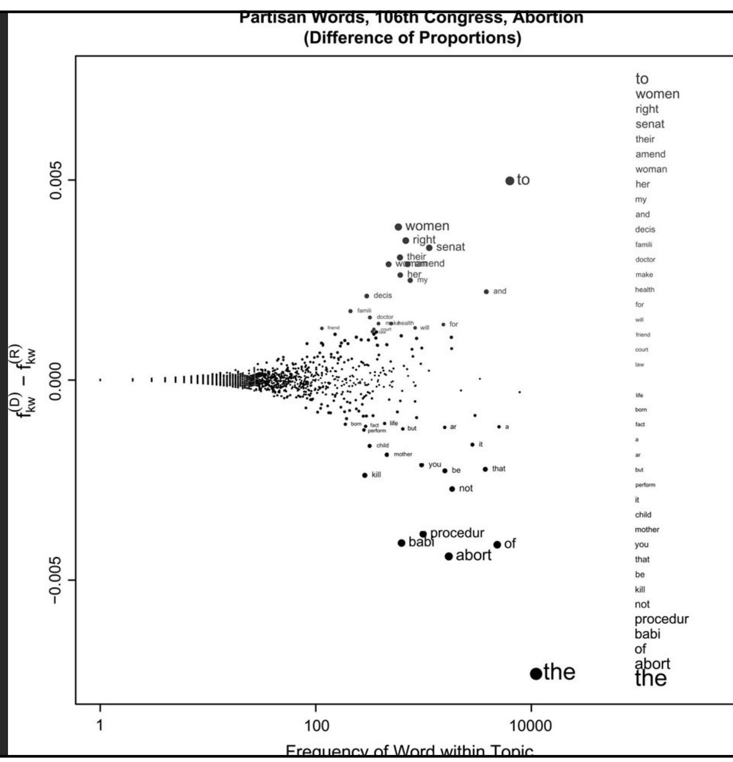
Term Frequency

tf_{td} = count(t) in d

Can take log frequency: $\log(1 + tf_{td})$

Can normalize to show proportion: $tf_{td} / \Sigma_t \, tf_{td}$

33



34

Keyword Weighting

Term Frequency

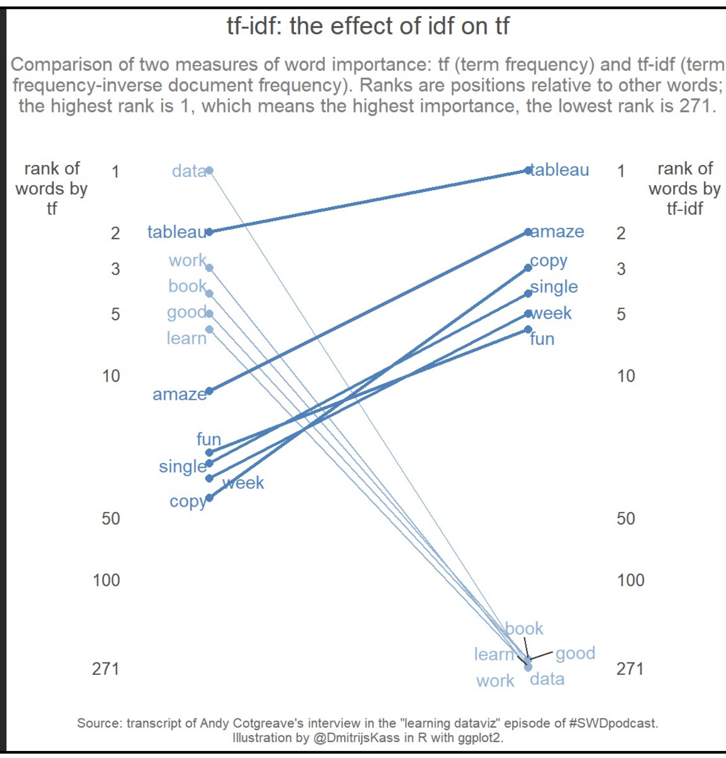
tf_{td} = count(t) in d

TF.IDF: Term Freq by Inverse Document Freq

$tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$

df_t = # docs containing t; N = # of docs

35



36

Limitations of Frequency Statistics

Typically focus on unigrams (single terms)

Often favors frequent (TF) or rare (IDF) terms

Not clear that these provide best description

"Bag of words" ignores additional info

Grammar / part-of-speech

Position within document

Recognizable entities

42

How do people describe text?

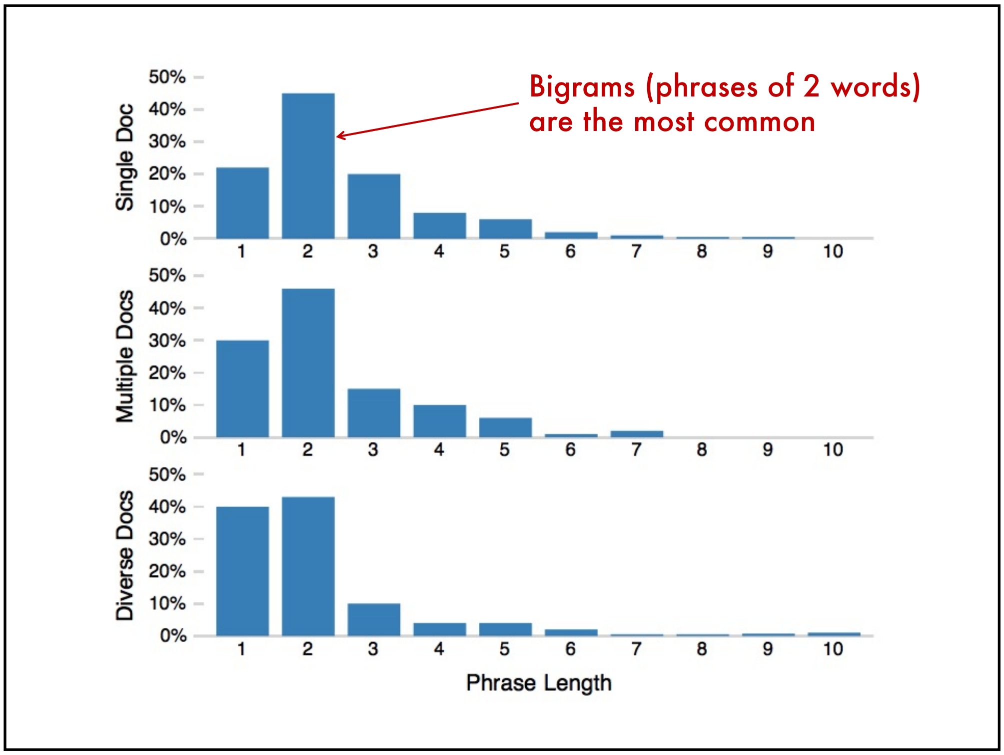
Asked 69 graduate students to read and describe dissertation abstracts

Each given 3 documents in sequence; summarized each using keyphrases, then summarized the 3 together as a whole using keyphrases

Were matched to both *familiar* and *unfamiliar* topics; *topical diversity* within a collection was varied systematically

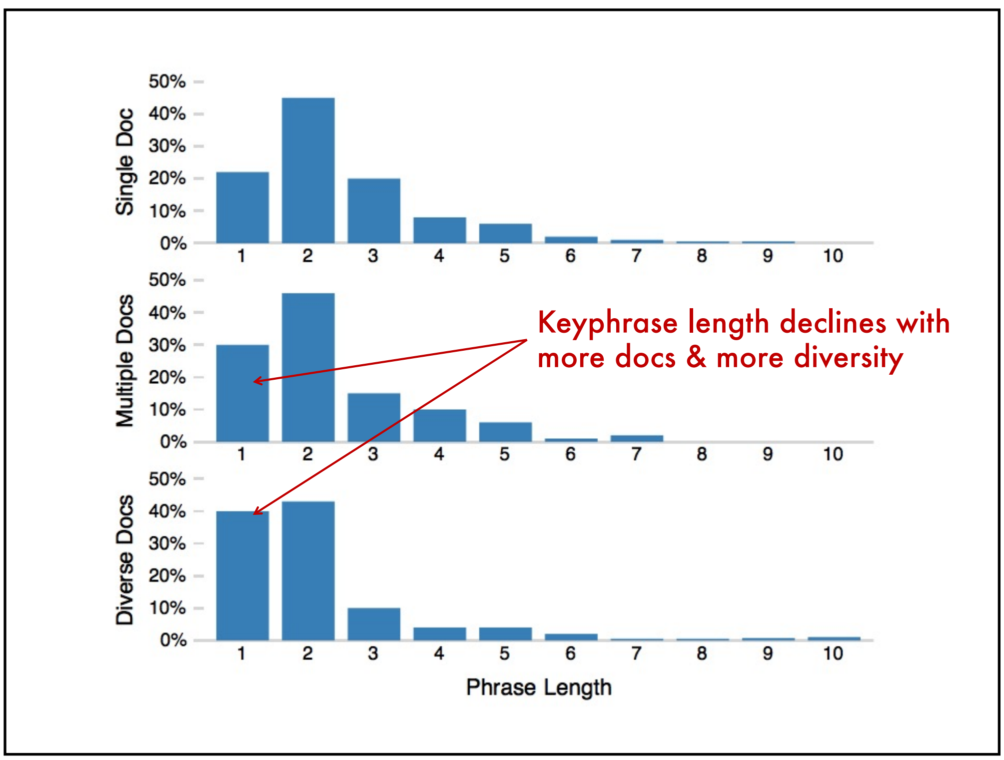
[Chuang 2012]

43



Bigrams (phrases of 2 words) are the most common
Single Doc
Multiple Docs
Diverse Docs
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
Phrase Length

44



Keyphrase length declines with more docs & more diversity
Single Doc
Multiple Docs
Diverse Docs
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
Phrase Length

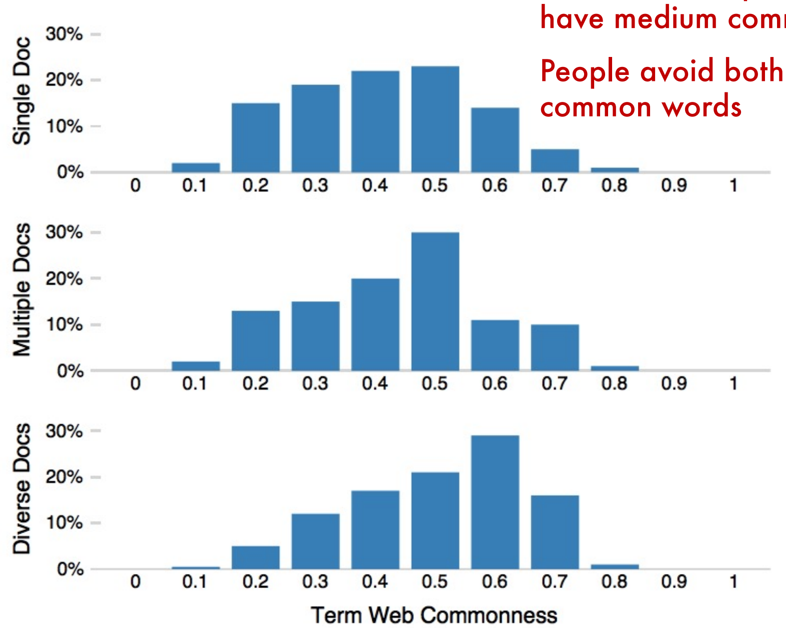
45

Term Commonness

$$\log(tf_w) \;/\; \log(tf_{the})$$

The normalized term frequency relative to the most frequent n-gram, e.g., the word "the".

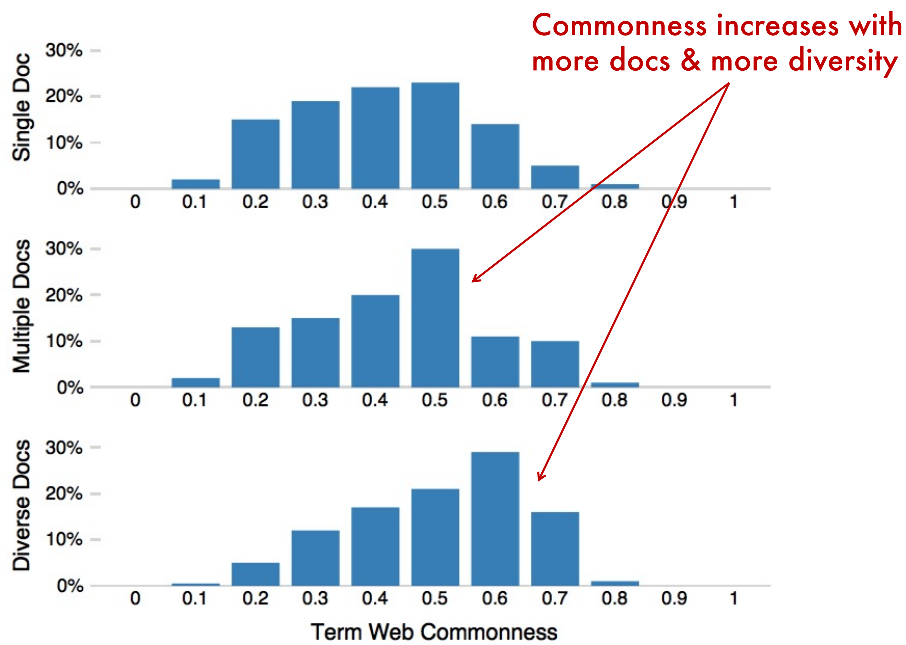
46



Selected descriptive terms have medium commonness

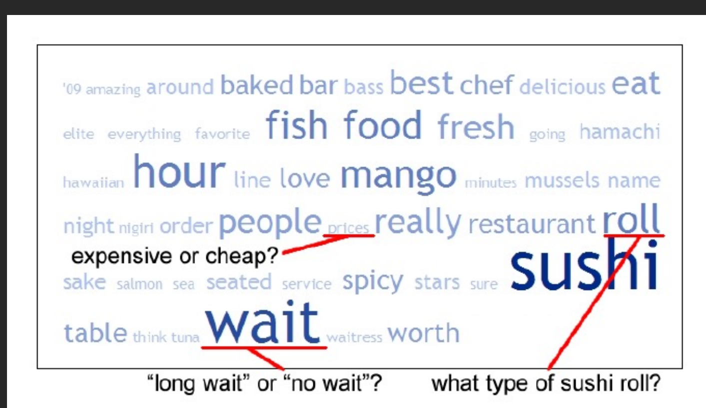
People avoid both rare and common words

47



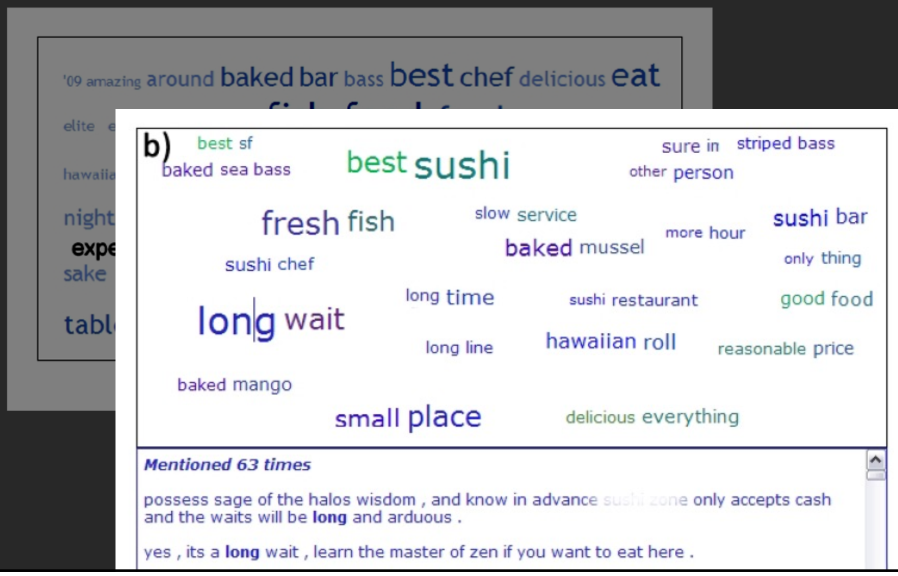
48

Yelp: Review Spotlight [Yatani 2011]



52

Yelp: Review Spotlight [Yatani 2011]



53

Tips: Descriptive Keyphrases

Understand the limitations of your language model

- **Bag of words:**
  - Easy to compute
  - Single words
  - Loss of word ordering

Select appropriate model and visualization

- Generate longer, more meaningful phrases
- Adjective-noun word pairs for reviews
- Show keyphrases within source text

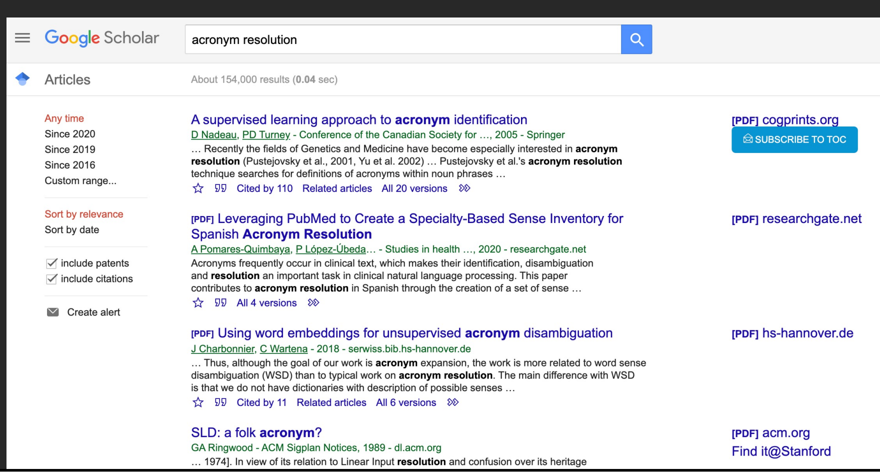
54

Visualizing Document Content

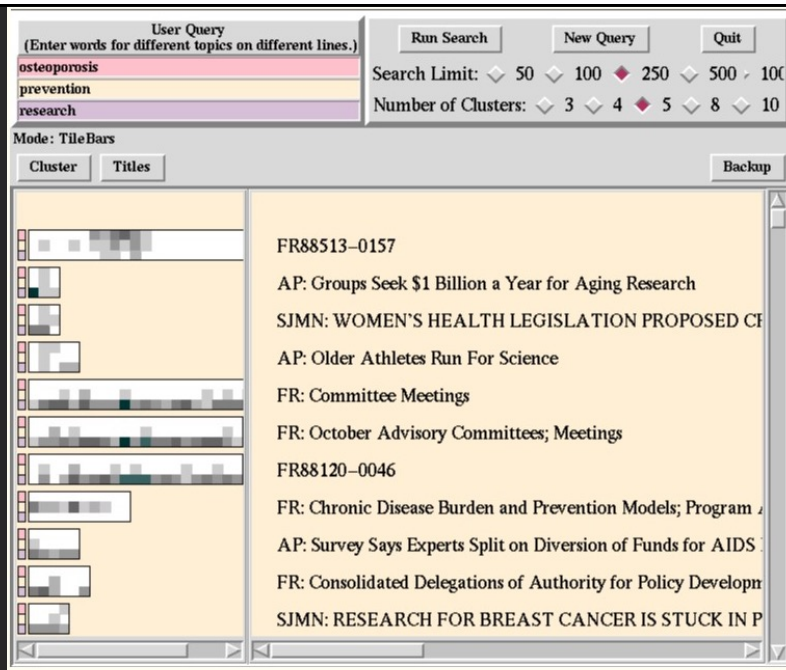
55

Information Retrieval

Search for documents
Match query string with documents
Visualization to **contextualize results**

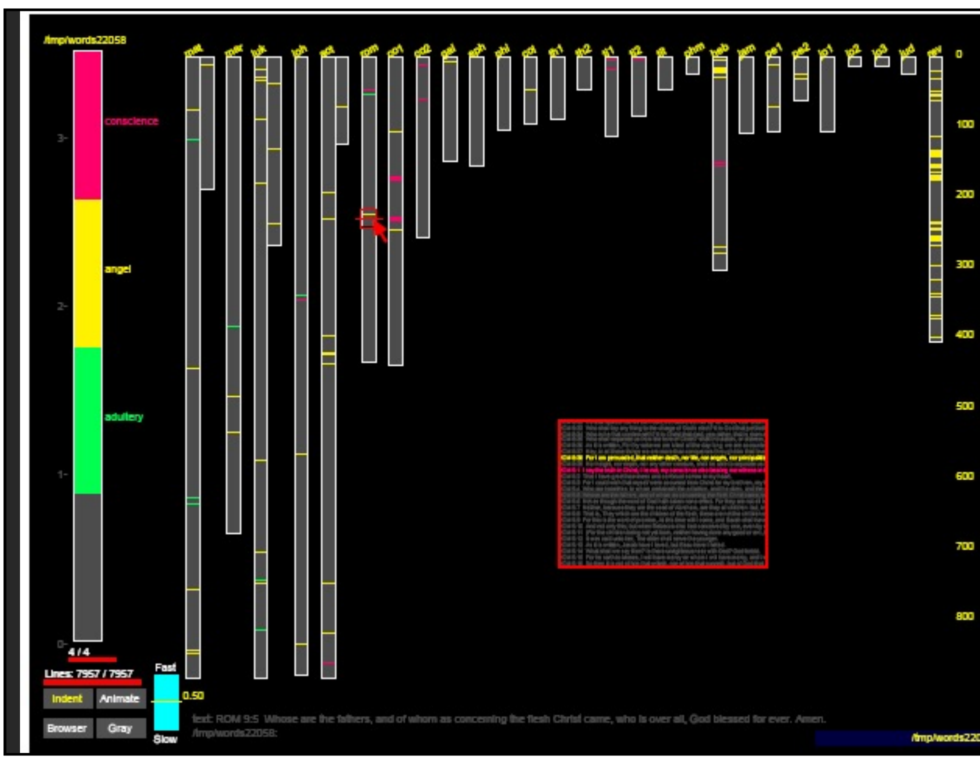


56

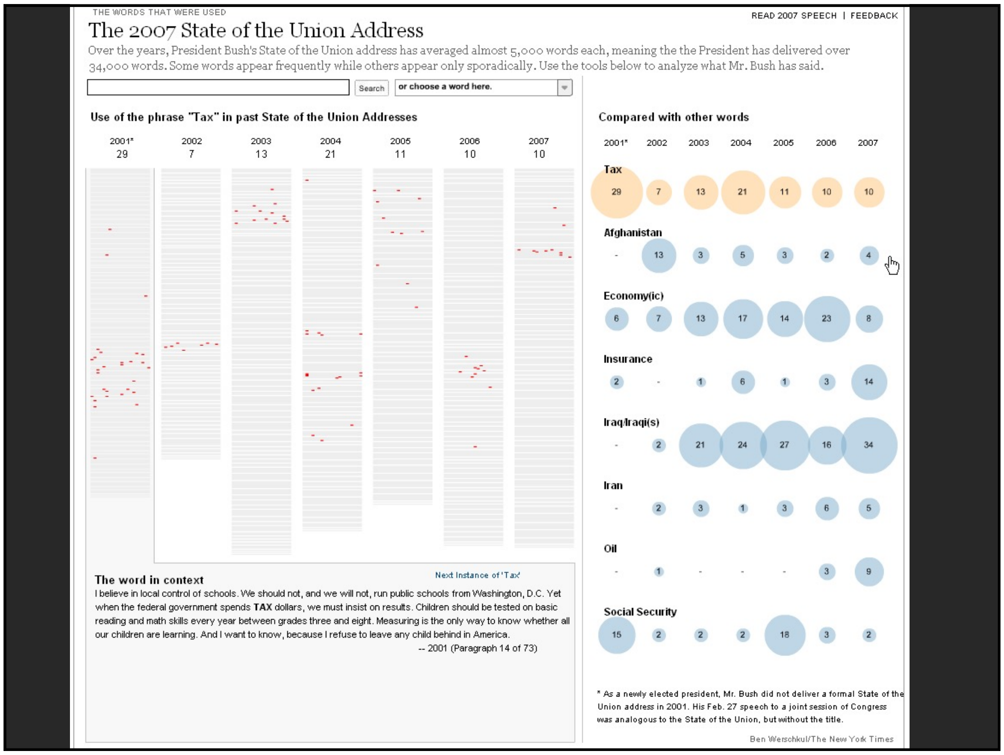


TileBars [Hearst]

57



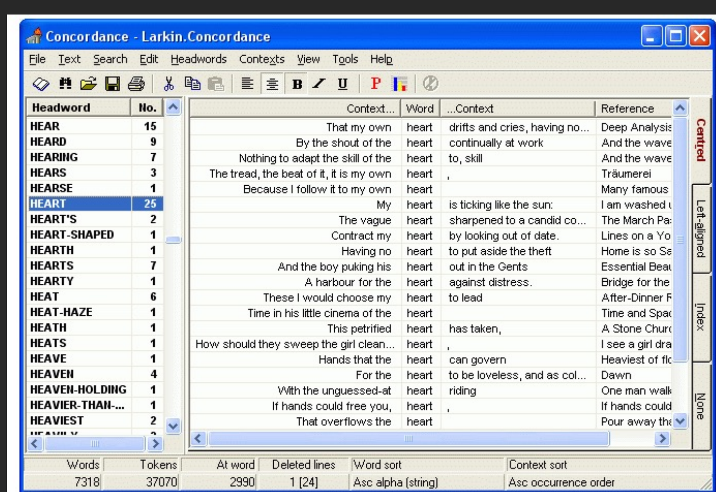
58



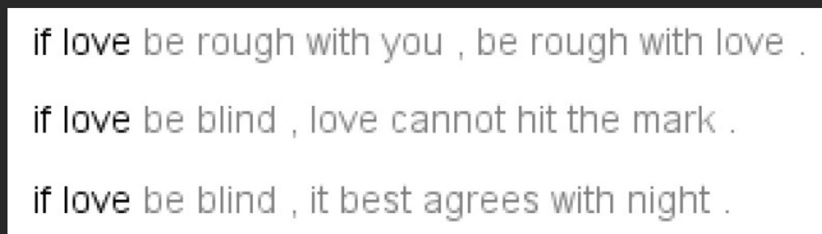
59

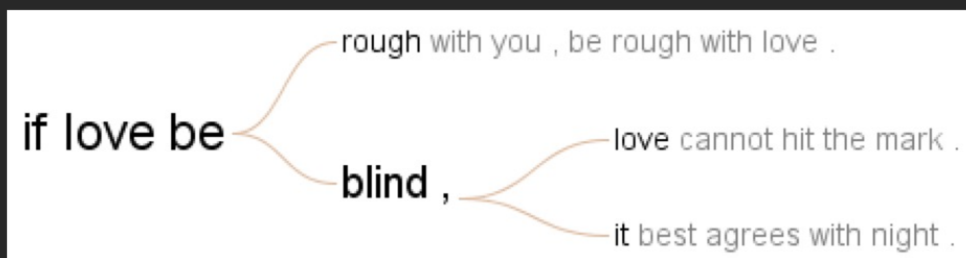
Concordance

What is the common local context of a term?



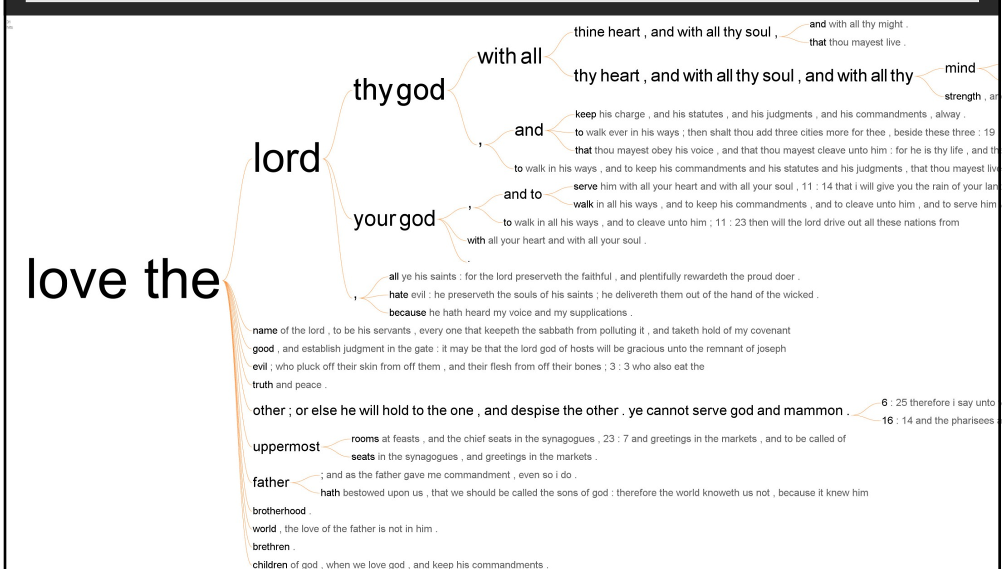
62





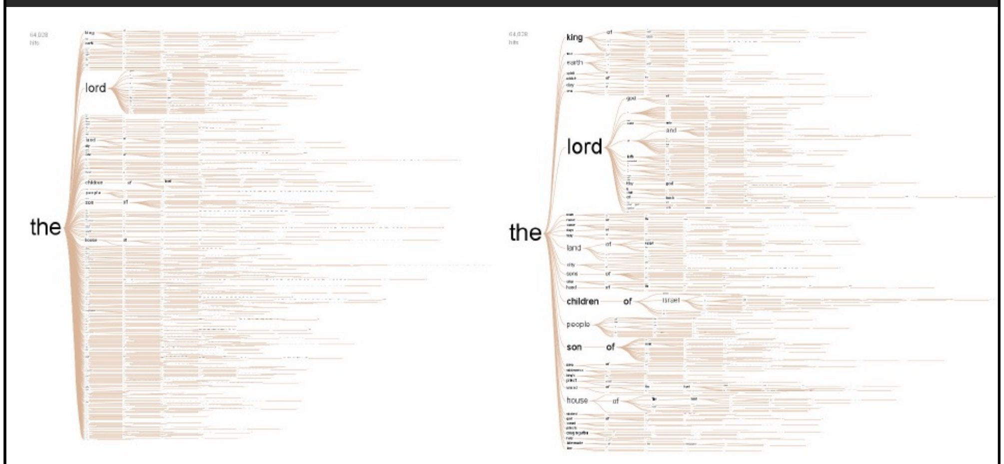
64

WordTree

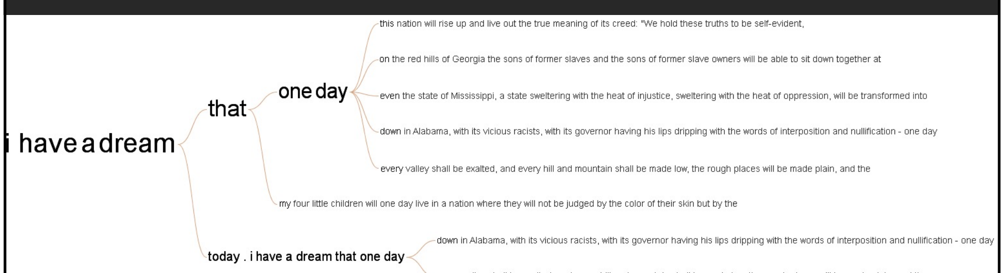


65

Filter infrequent runs

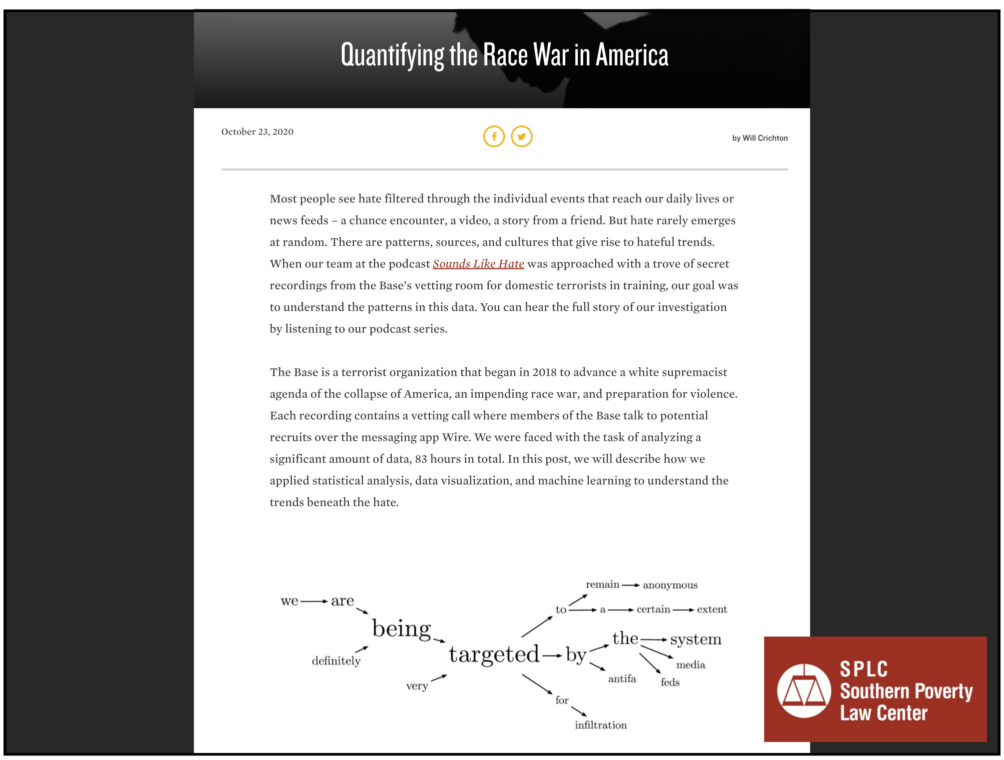


Recurrent themes in speech




word tree
we
must
act
will
are
, the people,
still believe that
have
cannot
all

68


Quantifying the Race War in America
October 23, 2020
by Will Crichton
Most people see hate filtered through the individual events that reach our daily lives or news feeds – a chance encounter, a video, a story from a friend. But hate rarely emerges at random. There are patterns, sources, and cultures that give rise to hateful trends. When our team at the podcast Sounds Like Hate was approached with a trove of secret recordings from the Base's vetting room for domestic terrorists in training, our goal was to understand the patterns in this data. You can hear the full story of our investigation by listening to our podcast series.
The Base is a terrorist organization that began in 2018 to advance a white supremacist agenda of the collapse of America, an impending race war, and preparation for violence. Each recording contains a vetting call where members of the Base talk to potential recruits over the messaging app Wire. We were faced with the task of analyzing a significant amount of data, 83 hours in total. In this post, we will describe how we applied statistical analysis, data visualization, and machine learning to understand the trends beneath the hate.
we
are
definitely
very
being
targeted
to
remain
anonymous
a
certain
extent
by
the
system
media
antifa
feds
for
infiltration
SPLC
Southern Poverty
Law Center

69

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>

70

Phrase Nets [van Ham 2009]

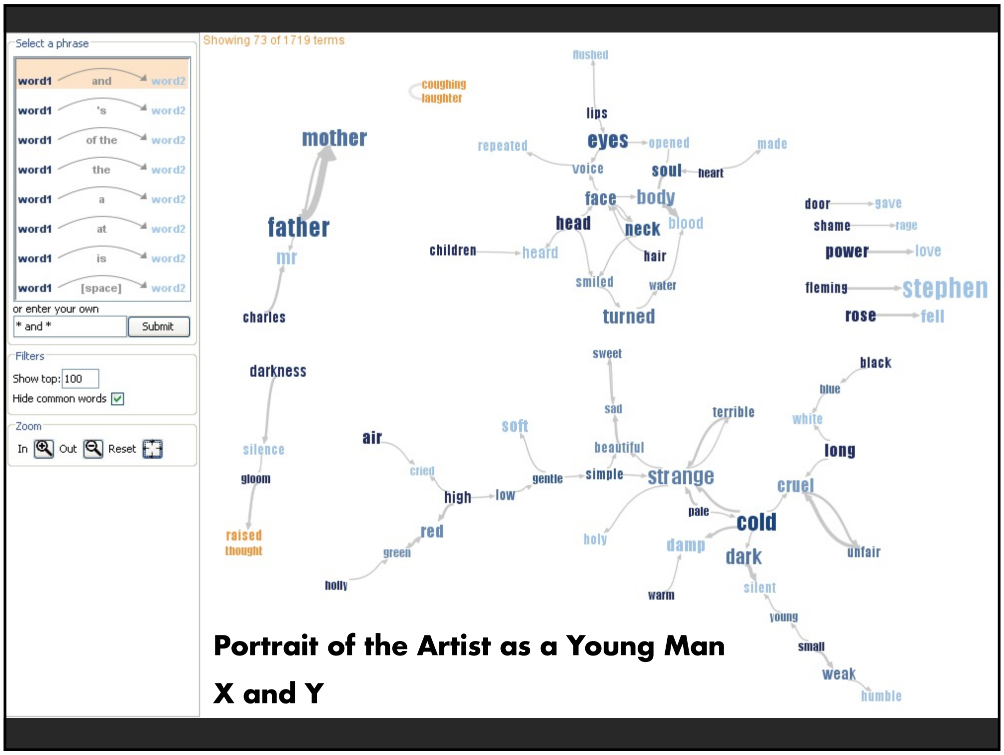
Look for specific linking patterns in the text:

- ‘A and B’, ‘A at B’, ‘A of B’, etc
- Could be output of regexp or parser

Visualize extracted patterns in a node-link view

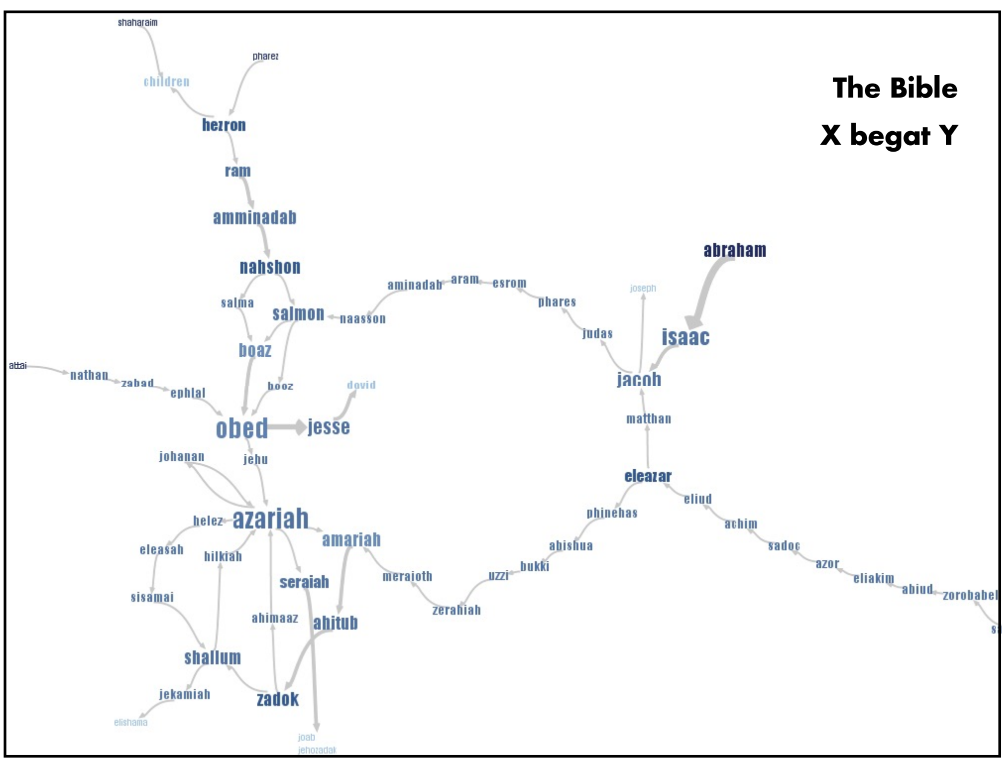
- Occurrences → Node size
- Pattern position → Edge direction
- Darker color → higher ratio of out-edges to in-edges

71



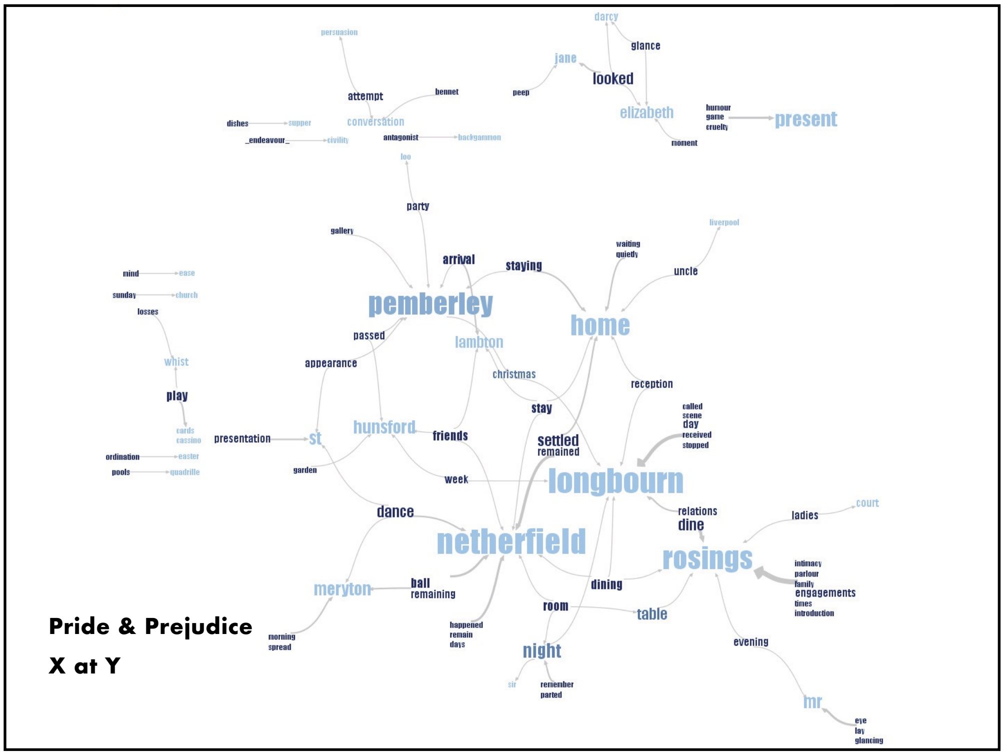
Select a phrase
word1 and word2
word1 's word2
word1 of the word2
word1 the word2
word1 a word2
word1 at word2
word1 is word2
word1 [space] word2
or enter your own
* and *
Submit
Filters
Show top: 100
Hide common words
Zoom
In Out Reset
Showing 73 of 1719 terms
flushed
coughing
laughter
lips
mother
repeated
eyes
opened
made
voice
soul
heart
father
face
body
head
neck
blood
door
gave
shame
rage
children
heard
hair
power
love
mr
smiled
water
fleming
stephen
charles
turned
rose
fell
sweet
darkness
black
sad
terrible
blue
soft
white
air
beautiful
long
silence
cried
gentle
simple
strange
cruel
gloom
high
low
pale
cold
red
holy
damp
raised
thought
green
dark
unfair
holly
silent
warm
young
small
weak
humble
Portrait of the Artist as a Young Man
X and Y

72

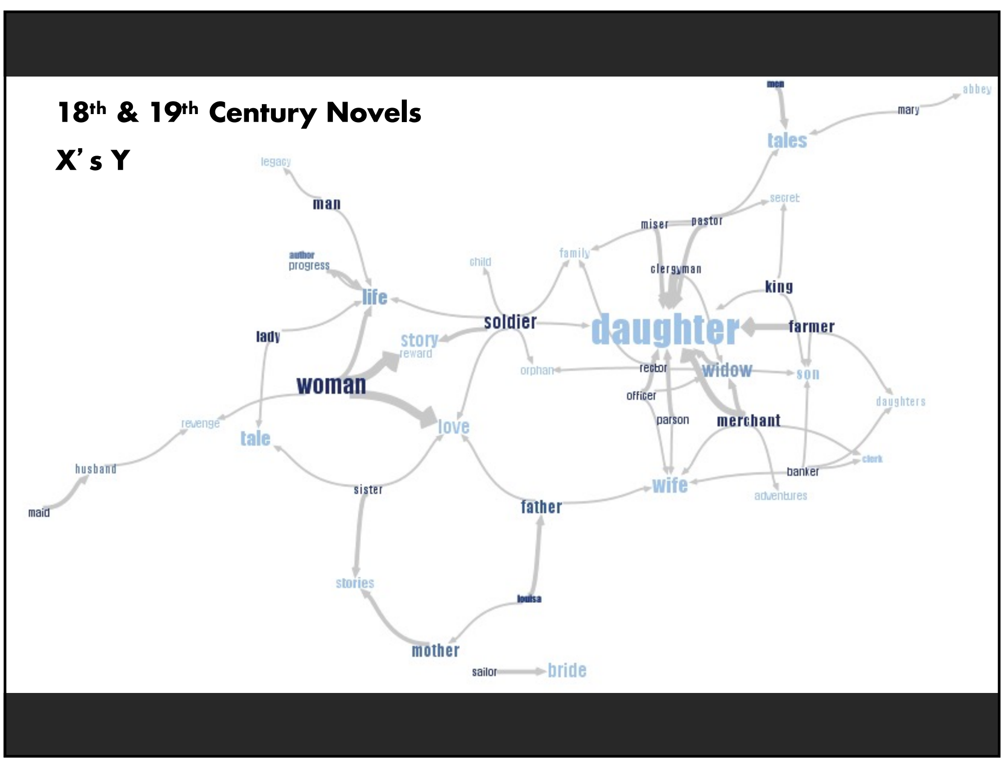


shaharaim
pharez
children
hezron
ram
amminadab
nahshon
abraham
aminadab
aram
esrom
joseph
salma
salmon
naasson
phares
judas
isaac
boaz
attai
nathan
zabad
ephlal
booz
david
jacob
obed
jesse
matthan
johanan
jehu
eleazar
eliud
phinehas
helez
azariah
achim
abishua
amariah
eleasah
hilkiah
sadoc
bukki
azor
seraiah
meraioth
uzzi
eliakim
abiud
zorobabel
sisamai
zerahiah
ahimaaz
ahitub
shallum
jekamiah
zadok
elishama
joab
jehozadak
The Bible
X begat Y

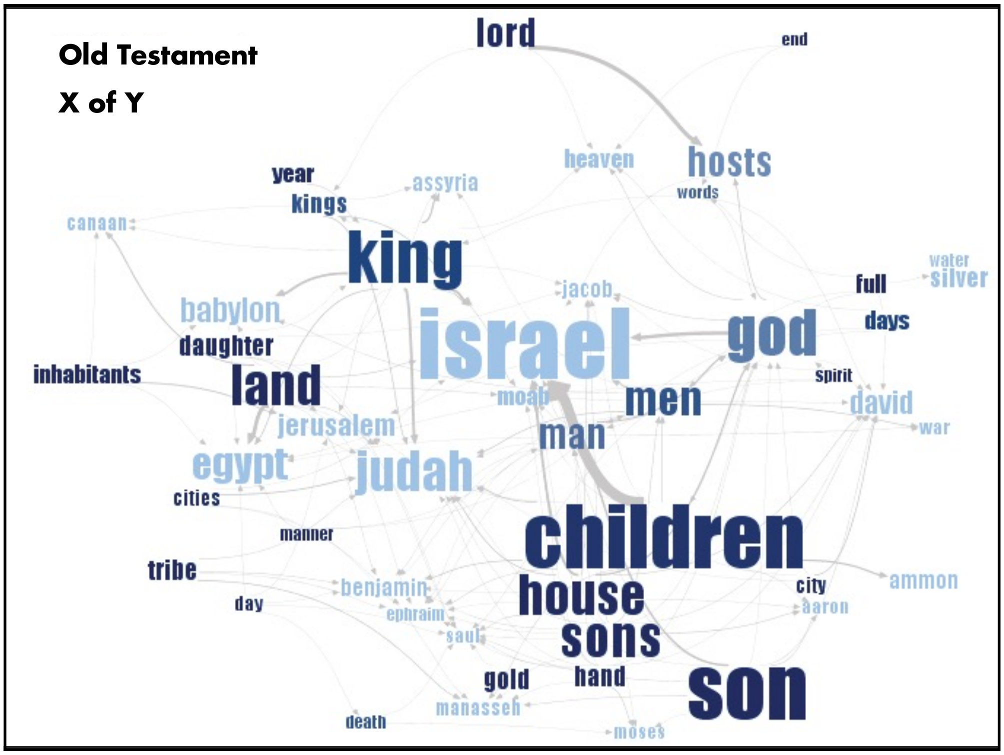
74


Pride & Prejudice
X at Y
darcy
glance
jane
looked
peep
elizabeth
present
persuasion
attempt
bennet
conversation
dishes
supper
endeavour
civility
antagonist
backgammon
loo
party
gallery
arrival
staying
waiting
quietly
uncle
liverpool
mind
ease
sunday
church
losses
pemberley
home
passed
lambton
appearance
whist
christmas
reception
play
stay
presentation
st
hunsford
friends
settled
remained
garden
week
longbourn
dance
relations
dine
ladies
court
netherfield
rosings
meryton
ball
remaining
dining
room
table
engagements
evening
night
mr

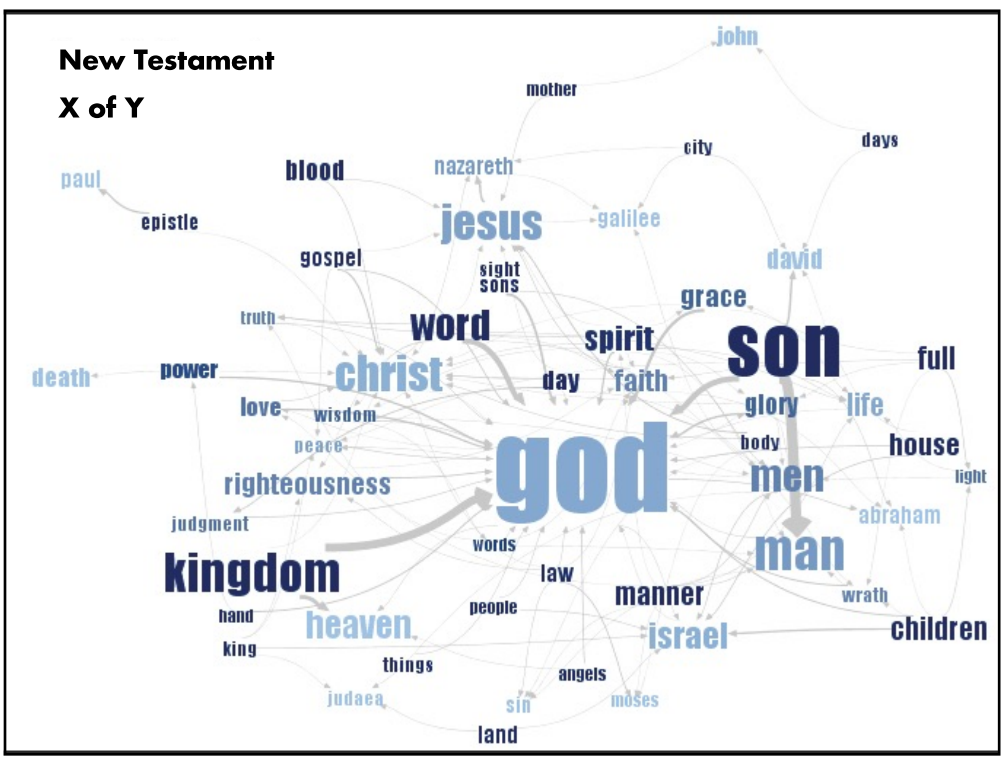
75


18th & 19th Century Novels
X's Y
men
abbey
mary
tales
legacy
man
secret
miser
pastor
author
progress
child
family
clergyman
life
king
lady
soldier
daughter
farmer
story
reward
orphan
rector
widow
son
woman
officer
daughters
parson
merchant
revenge
love
tale
clerk
husband
banker
sister
wife
maid
adventures
father
stories
louisa
mother
sailor
bride

78


Old Testament
X of Y
lord
end
hosts
heaven
words
year
kings
assyria
canaan
king
jacob
full
water
silver
babylon
israel
god
days
daughter
inhabitants
land
spirit
moab
men
david
jerusalem
man
war
egypt
judah
cities
manner
children
tribe
benjamin
house
city
ammon
day
ephraim
aaron
saul
sons
gold
hand
son
manasseh
death
moses

79


New Testament
X of Y
john
mother
city
days
paul
blood
nazareth
epistle
jesus
galilee
gospel
david
sight
sons
grace
truth
word
spirit
son
full
death
power
christ
day
faith
love
wisdom
glory
life
peace
body
house
god
men
light
righteousness
abraham
judgment
man
words
kingdom
law
wrath
manner
hand
heaven
people
israel
children
king
things
angels
judaea
sin
moses
land

80

Visualizing Conversation

91

Visualizing Conversation

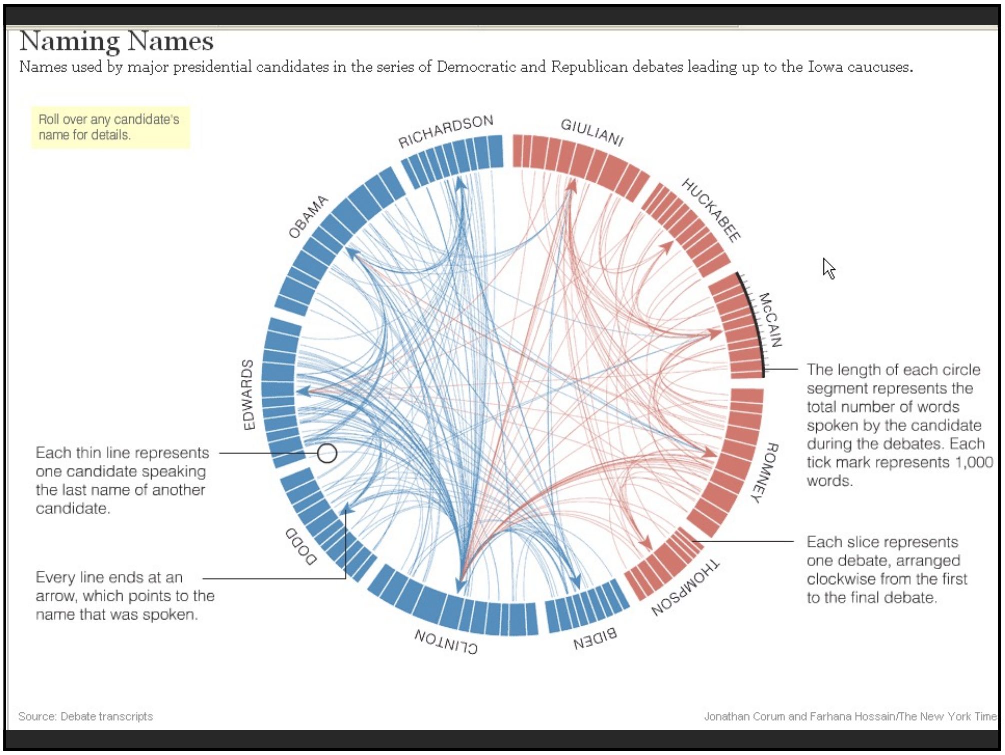
Many dimensions to consider:

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

Interesting cross-products:

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

92

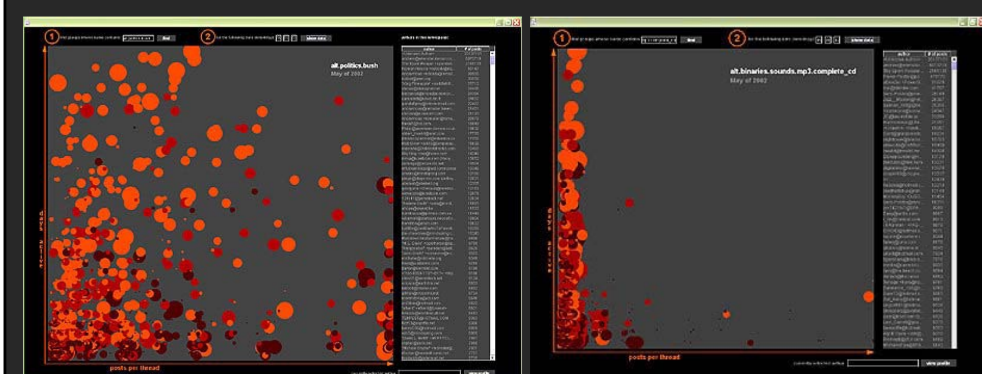


93

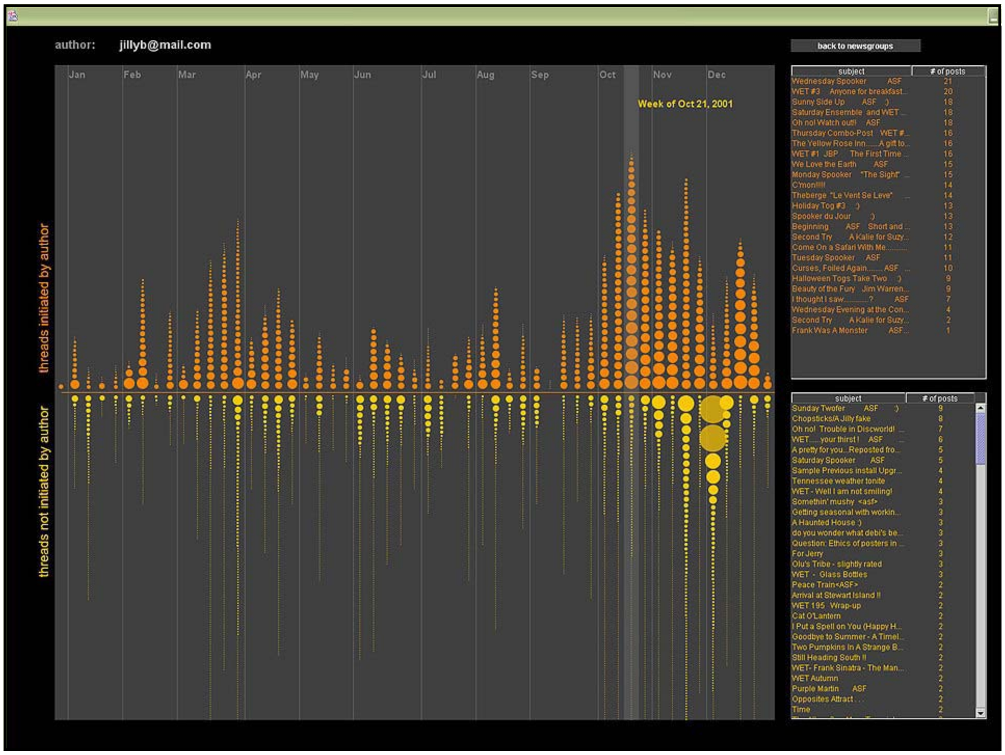
Usenet Visualization [Viégas]

Show correspondence patterns in text forums

Initiate vs. reply; size and duration of discussion

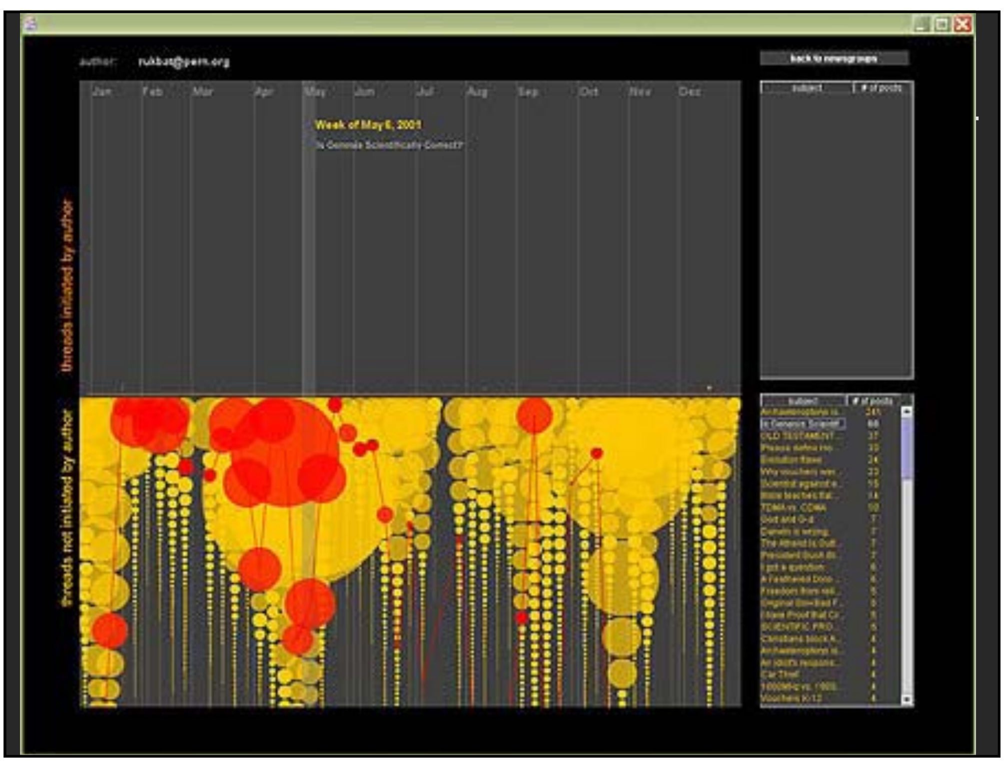


96



author: jillyb@mail.com
back to newsgroups
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Week of Oct 21, 2001
threads initiated by author
threads not initiated by author
subject # of posts
Wednesday Spooker ASF 21
WET #3 Anyone for breakfast... 20
Sunny Side Up ASF :) 18
Saturday Ensemble and WET ... 18
Oh no! Watch out! ASF 18
Thursday Combo-Post WET #... 16
The Yellow Rose Inn.......A gift to... 16
WET #1 JBP The First Time ... 16
We Love the Earth ASF 15
Monday Spooker "The Sight" ... 15
C'mon!!!! 14
Theberge "Le Vent Se Leve" ... 14
Holiday Tog #3 :) 13
Spooker du Jour :) 13
Beginning ASF Short and ... 13
Second Try A Katie for Suzy... 12
Come On a Safari With Me.......... 11
Tuesday Spooker ASF 11
Curses, Foiled Again........ ASF ... 10
Halloween Togs Take Two :) 9
Beauty of the Fury Jim Warren... 9
I thought I saw............? ASF 7
Wednesday Evening at the Con... 4
Second Try A Katie for Suzy... 2
Frank Was A Monster ASF... 1
subject # of posts
Sunday Twofer ASF :) 9
Chopsticks/A Jilly fake 8
Oh no! Trouble in Discworld! ... 7
WET......your thirst ! ASF ... 6
A pretty for you...Reposted fro... 5
Saturday Spooker ASF 5
Sample Previous install Upgr... 4
Tennessee weather tonite 4
WET - Well I am not smiling! 4
Somethin' mushy <asf> 3
Getting seasonal with workin... 3
A Haunted House :) 3
do you wonder what debi's be... 3
Question: Ethics of posters in ... 3
For Jerry 3
Olu's Tribe - slightly rated 3
WET - Glass Bottles 3
Peace Train<ASF> 2
Arrival at Stewart Island !! 2
WET 195 Wrap-up 2
Cat O'Lantern 2
I Put a Spell on You (Happy H... 2
Goodbye to Summer - A Timel... 2
Two Pumpkins In A Strange B... 2
Still Heading South !! 2
WET- Frank Sinatra - The Man... 2
WET Autumn 2
Purple Martin ASF 2
Opposites Attract . . . 2
Time 2

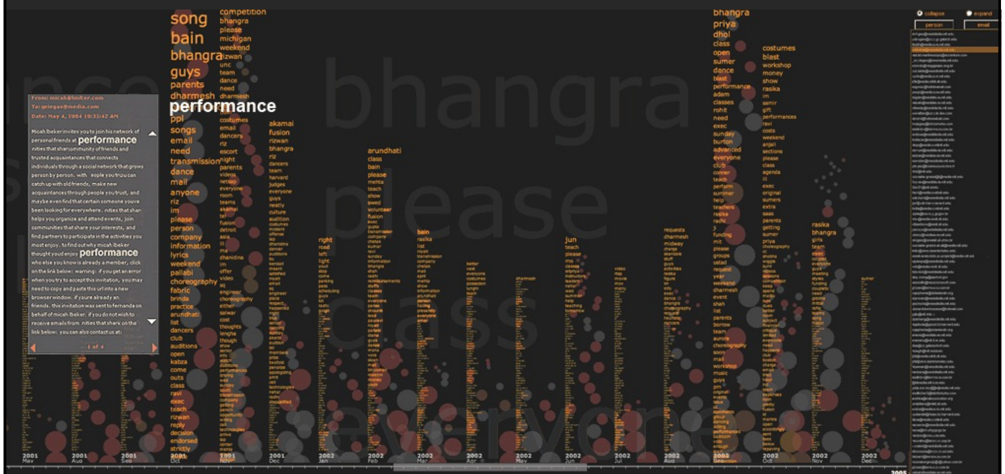
97



author: rukbat@pern.org
back to newsgroups
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Week of May 6, 2001
Is Genesis Scientifically Correct?
threads initiated by author
threads not initiated by author
subject # of posts

98

Themail (Viégas)



One person over time, TF.IDF weighted terms

100

Visualizing Document Collections

104

Topic modeling

Topic modeling approaches

Assume documents are a mixture of topics

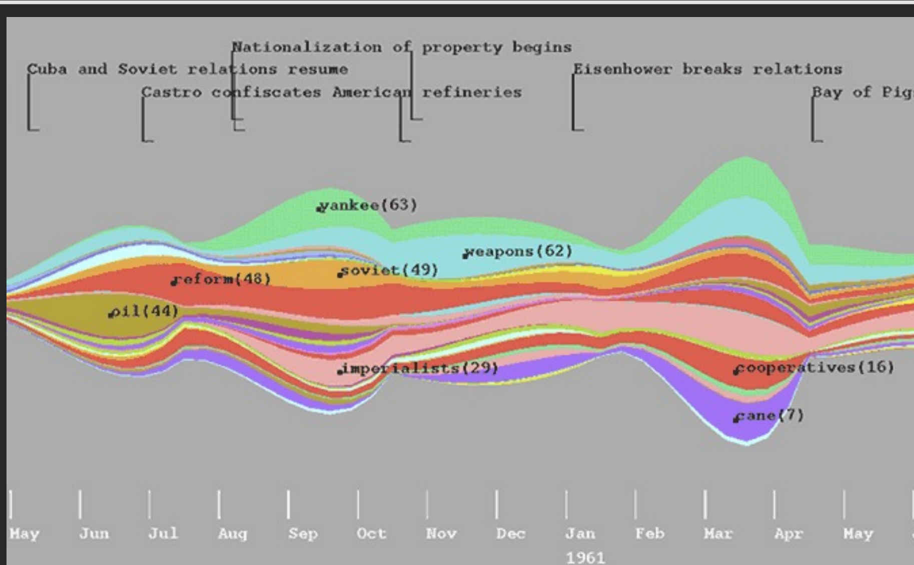
Topics are (roughly) a set of co-occurring terms

Latent Semantic Analysis (LSA): reduce term matrix

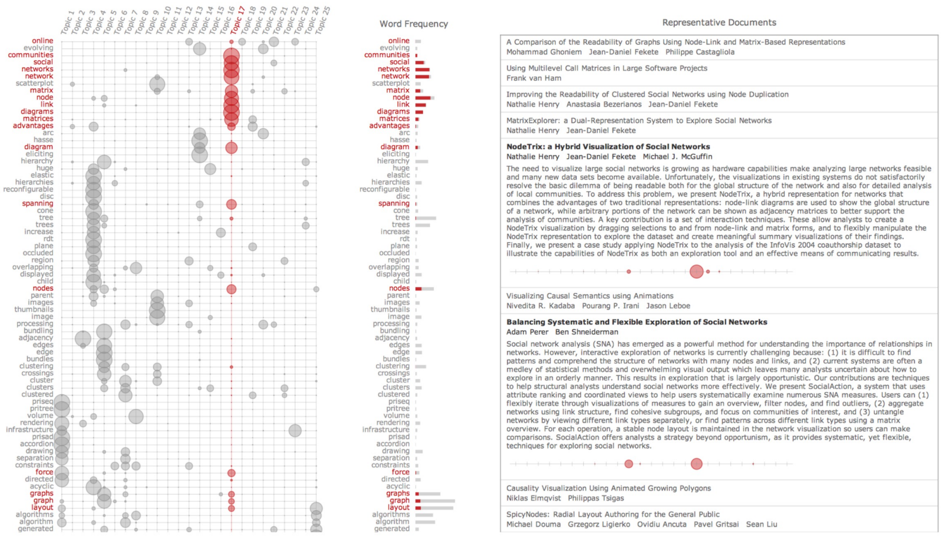
Latent Dirichlet Allocation (LDA): statistical model

113

ThemeRiver (Havre et al 99)



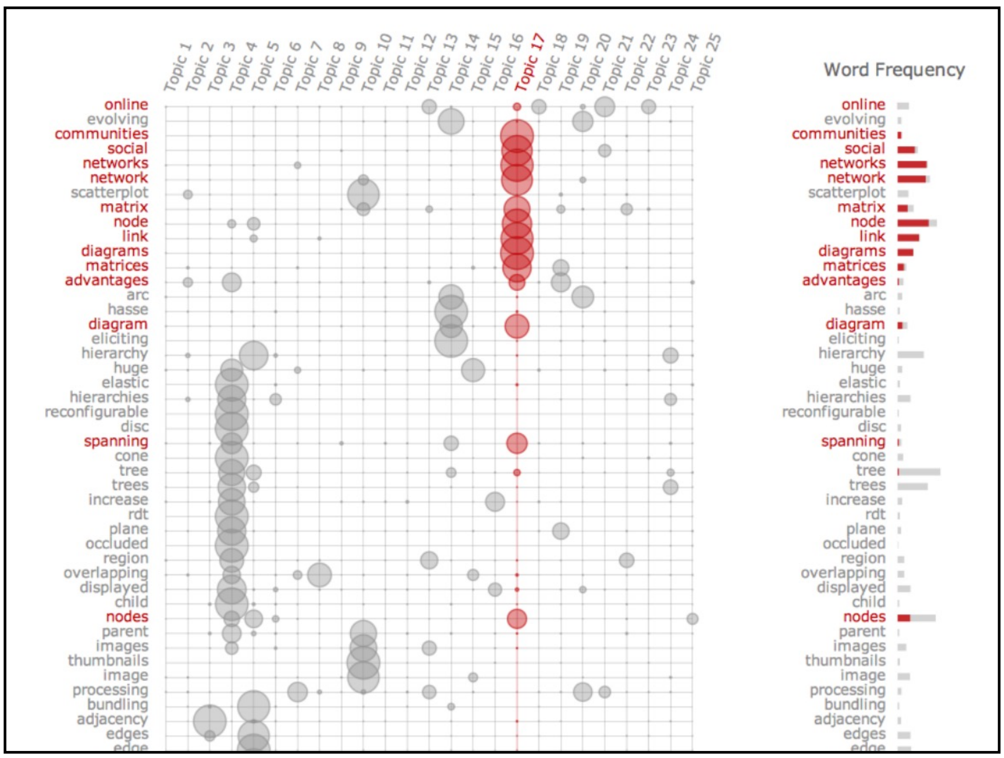
114



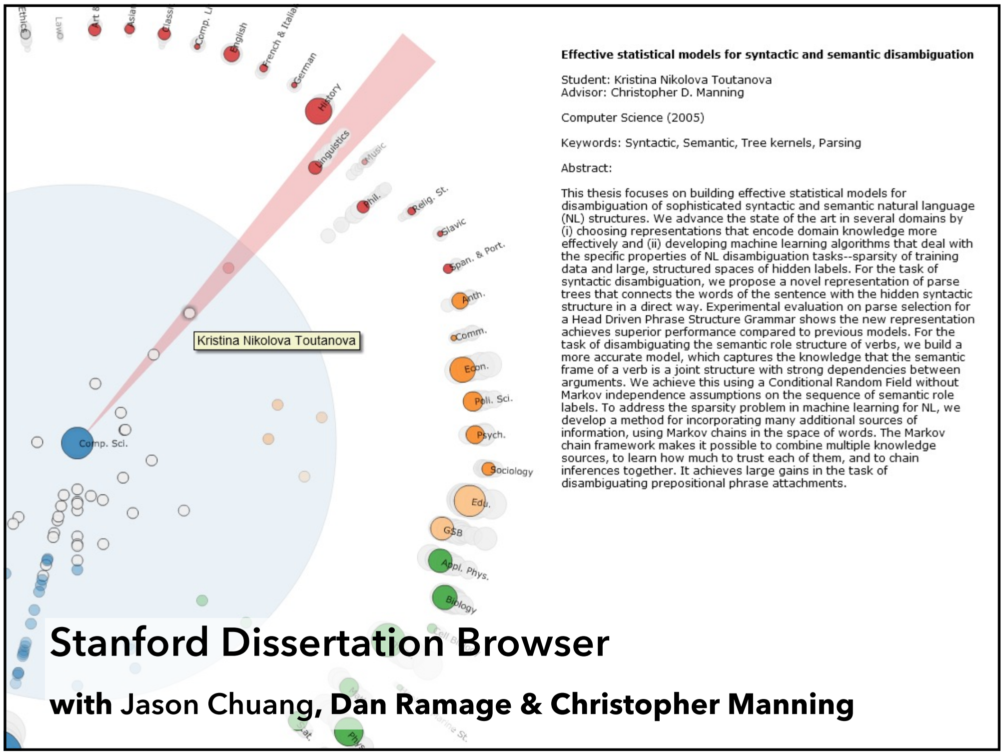
Termite: Visualizing Topic Models [Chuang '12]

Show salient (vs. frequent) terms. Seriate rows & columns.

122

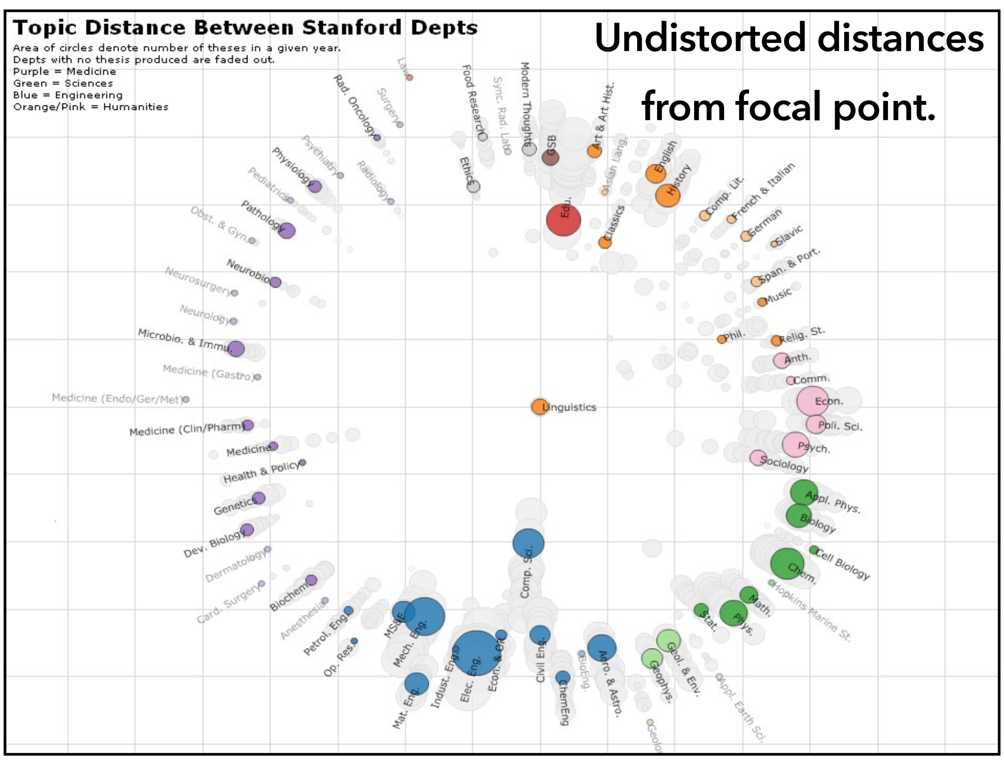


123



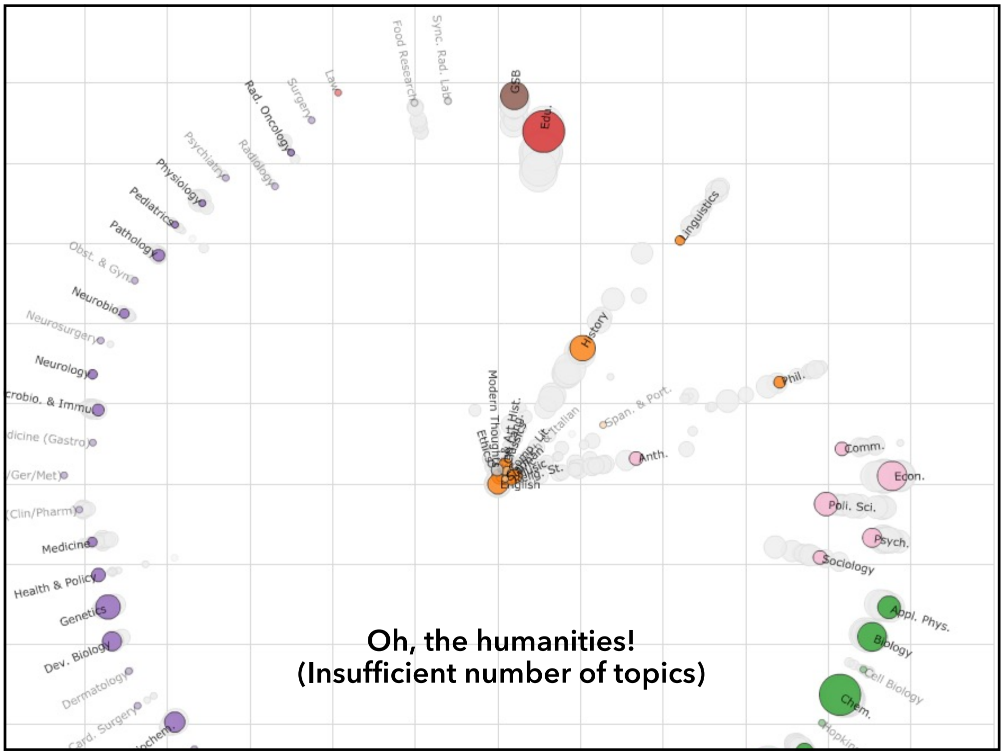
Effective statistical models for syntactic and semantic disambiguation
Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning
Computer Science (2005)
Keywords: Syntactic, Semantic, Tree kernels, Parsing
Abstract:
This thesis focuses on building effective statistical models for disambiguation 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.
Kristina Nikolova Toutanova
Comp. Sci.
History
Linguistics
Phil.
Relig. St.
Slavic
Span. & Port.
Anth.
Comm.
Econ.
Poli. Sci.
Psych.
Sociology
Edu.
GSB
Appl. Phys.
Biology
Stanford Dissertation Browser
with Jason Chuang, Dan Ramage & Christopher Manning

124



Topic Distance Between Stanford Depts
Area of circles denote number of theses in a given year.
Depts with no thesis produced are faded out.
Purple = Medicine
Green = Sciences
Blue = Engineering
Orange/Pink = Humanities
Undistorted distances
from focal point.
Linguistics
Edu.
GSB
Art & Art Hist.
English
History
Classics
Comp. Lit.
French & Italian
German
Slavic
Span. & Port.
Music
Phil.
Relig. St.
Anth.
Comm.
Econ.
Poli. Sci.
Psych.
Sociology
Appl. Phys.
Biology
Cell Biology
Chem.
Math.
Phys.
Stat.
Geol. & Env.
Geophys.
Appl. Earth Sci.
Aero. & Astro.
ChemEng
Civil Eng.
Comp. Sci.
Econ. & Ort.
Elec. Eng.
Indust. Eng.
Mat. Eng.
Mech. Eng.
MS&E
Op. Res.
Petrol. Eng.
Biochem.
Dev. Biology
Genetics
Health & Policy
Medicine
Medicine (Clin/Pharm)
Medicine (Endo/Ger/Met)
Medicine (Gastro)
Microbio. & Immu.
Neurology
Neurosurgery
Neurobio.
Obst. & Gyn.
Pathology
Pediatrics
Physiology
Psychiatry
Radiology
Rad. Oncology
Surgery
Food Research
Ethics
Modern Thought

126



127

Summary

High Dimensionality

Where possible use text to represent text…
… which terms are the most descriptive?

Context & Semantics

Provide relevant context to aid understanding.
Show (or provide access to) the source text.

134