

Text

C. Andrews

2016-05-09

Why visualize text?

Visualization goals

Understanding - read a document

Summaries - get the "gist" of a document

Clustering - group together similar contents

Correlate - compare patterns in text to other data

High-level tasks

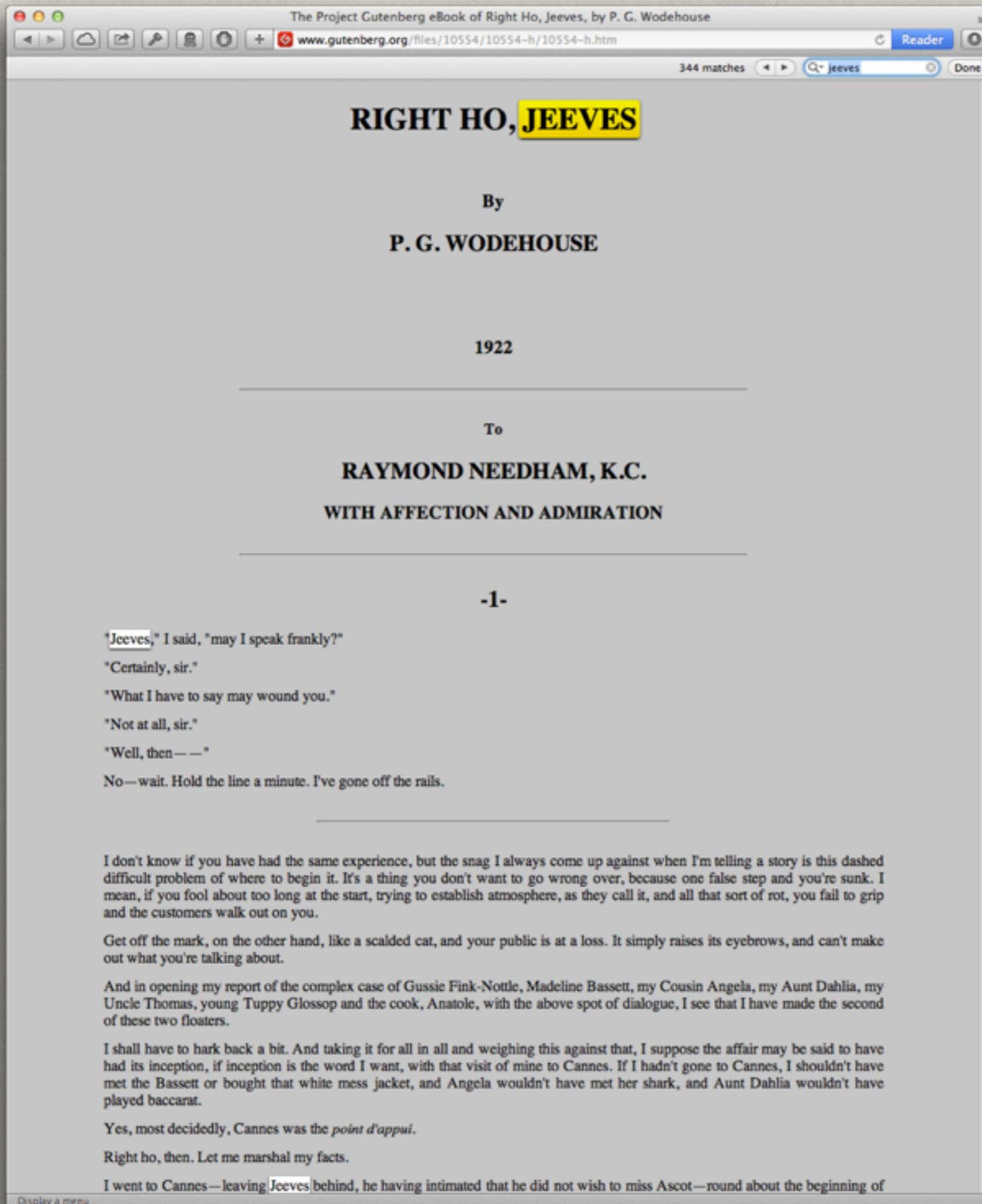
Find documents of interest in a collection

Find documents similar to ones I want

Identify the main themes or ideas of documents or collections

See the history of changes in a document

Find connections between documents



Interaction tool includes context of position and length of document

All details available through interaction

Search is built in

Tool is familiar, readily available and convenient

```
census1.html UNREGISTERED  
census1.html  
32  
33 function makeScatterplot(){  
34   var margin = {top:20, bottom:20, left:60, right: 20};  
35   var width = 500, height = 500;  
36   var xValue = function(d){return d[0]};  
37   var yValue = function(d){return d[1]};  
38   var xScale = d3.scale.linear();  
39   var yScale = d3.scale.linear();  
40   var xAxis = d3.svg.axis().scale(xScale).orient("bottom");  
41   var yAxis = d3.svg.axis().scale(yScale).orient("left");  
42  
43  
44  
45 function chart(selection){  
46   selection.each(function(data){  
47     xScale.range([0,width - margin.left - margin.right])  
48     .nice()  
49     .domain(d3.extent(data, xValue));  
50  
51     yScale.range([height - margin.top - margin.bottom, 0])  
52     .nice()  
53     .domain(d3.extent(data, yValue));  
54  
55     var svg = d3.select(this).append("svg")  
56     .attr({width:width, height:height});  
57  
58     var canvas = svg.append("g")  
59     .attr("transform","translate("+margin.left +","+margin.top+)");  
60  
61     // create the dots  
62     var dots = canvas.selectAll("circle")  
63     .data(data)  
64     .enter()  
65     .append("circle");  
66  
67  
68
```



Line 1, Column 16 Tab Size: 4 HTML

overview + detail display

syntactic structure is mapped to a color encoding

I don't know if you have had the same experience, but the snag I always come up against when I'm telling a story is this dashed difficult problem of where to begin it. It's a thing you don't want to go wrong over, because one false step and you're sunk. I mean, if you fool about too long at the start, trying to establish atmosphere, as they call it, and all that sort of rot, you fail to grip and the customers walk out on you.

Get off the mark, on the other hand, like a scalded cat, and your public is at a loss. It simply raises its eyebrows, and can't make out what you're talking about.

And in opening my report of the complex case of Gussie Fink-Nottle, Madeline Bassett, my Cousin Angela, my Aunt Dahlia, my Uncle Thomas, young Tuppy Glossop and the cook, Anatole, with the above spot of dialogue, I see that I have made the second of these two floaters.

I shall have to hark back a bit. And taking it for all in all and weighing this against that, I suppose the affair may be said to have had its inception, if inception is the word I want, with that visit of mine to Cannes. If I hadn't gone to Cannes, I shouldn't have met the Bassett or bought that white mess jacket, and Angela wouldn't have met her shark, and Aunt Dahlia wouldn't have played baccarat.

Yes, most decidedly, Cannes was the *point d'appui*.

Right ho, then. Let me marshal my facts.

LO, praise of the prowess of people-kings
of spear-armed Danes, in days long sped,
we have heard, and what honor the athelings won!
Oft Scyld the Scefing from squadroned foes,
from many a tribe, the mead-bench tore,
awing the earls. Since erst he lay
friendless, a foundling, fate repaid him:
for he waxed under welkin, in wealth he throve,
till before him the folk, both far and near,
who house by the whale-path, heard his mandate,
gave him gifts: a good king he!
To him an heir was afterward born,
a son in his halls, whom heaven sent
to favor the folk, feeling their woe
that erst they had lacked an earl for leader
so long a while; the Lord endowed him,
the Wielder of Wonder, with world's renown.
Famed was this Beowulf: far flew the boast of him,
son of Scyld, in the Scandian lands.
So becomes it a youth to quit him well
with his father's friends, by fee and gift,
that to aid him, aged, in after days,
come warriors willing, should war draw nigh,
liegemen loyal: by lauded deeds
shall an earl have honor in every clan.

YOU don't know about me without you have read a book by the name of The Adventures of Tom Sawyer; but that ain't no matter. That book was made by Mr. Mark Twain, and he told the truth, mainly. There was things which he stretched, but mainly he told the truth. That is nothing. I never seen anybody but lied one time or another, without it was Aunt Polly, or the widow, or maybe Mary. Aunt Polly—Tom's Aunt Polly, she is—and Mary, and the Widow Douglas is all told about in that book, which is mostly a true book, with some stretchers, as I said before.

1. US GOVERNMENT INTELLIGENCE REPORT: 7 OCTOBER 2008 from MI-5

MI-5 states that 100,000 bullets, six guns and military ware were among items police say they recovered after raiding a house in Narok, Kenya, belonging to Mr Thabiti Otieno. The MI5 officers are investigating how sleeping bags, Nato-approved 7.62mm calibre ammunition, military desert boots, combat rain jackets, military fuel tanks, British Army inscribed Land Rover spare parts, military machetes and knives, landed in the hands of an unauthorised person.

3. US GOVERNMENT TELEPHONE INTERCEPT: [October 10 2008]

Call placed from a pre-paid cell phone (Caller 1) in an apartment complex in Barcelona, Venezuela to a pre-paid cell phone (Caller 2) in a business district Carabobo, Venezuela. Conversation took place in Spanish at about 1217, Barcelona local time.

Caller 1: Jorge's place was raided last night. I think they [probably the police] were watching the place. They found the guns-

Caller 2: (interrupted) Shut up! Shut up! Just shut up! You can't say that.

Caller 1: Sorry, and we need a new source for the, uh, car parts.

Caller 2: I think I have a source. My friend has a connection to a guy who is selling parts. I'm going to contact him today or tomorrow.

Caller 1: When do you think we can get them?

Caller 2: I don't know. Just don't talk to anybody. I'll call you on this phone when we have anything. Don't call me and don't do anything.

[END]

1. News Article: Bangkok Post, March 2, 2008

Thai authorities seized an Ilyushin IL-76 aircraft carrying tons of weapons from North Korea during a refueling stop in Bangkok, a government official said. The pilot told Thai authorities the aircraft was headed to Sri Lanka, but its final destination was unknown, according to a spokesman for the Thai prime minister. It contained about 35 tons of weapons, including rocket-propelled grenades, shoulder-launched rockets and tubes that may be missile components, the spokesman said. The plane, which was detained Saturday, had five people onboard -- four from Ukraine and one from Belarus. They will appear in court Monday on charges related to illegal weapons smuggling, the spokesman said.

1. REPORT DATE: 9 October 2008 [provided to CIA by Pakistani Criminal Investigation Unit, Karachi Division]

NOTE: Surveillance report on the activities of Maulana Haq Bukhari, suspected to be a top leader within the Karachi faction of Lashkar-e-Jhangvi. Bukhari is frequently accompanied by Akram Basra, who acts as a driver and bodyguard. Additional information provided by police informants.

23 July 2008 – A delivery was made to a house in Lyari Town (a constituent town of Karachi) in a house believed to be used by Bukhari. The delivery was made by a two men in street clothing (as opposed to a uniform) who arrived in a white van, license LHR 6354, with single blue stripe on each side. The delivery consisted of three medium boxes (requiring two hands to move) and a small box (handsized). The boxes appeared to be heavy. One large box was square, the other rectangular. The small box was rectangular. It is unknown if Bukhari was home at the time of the delivery.

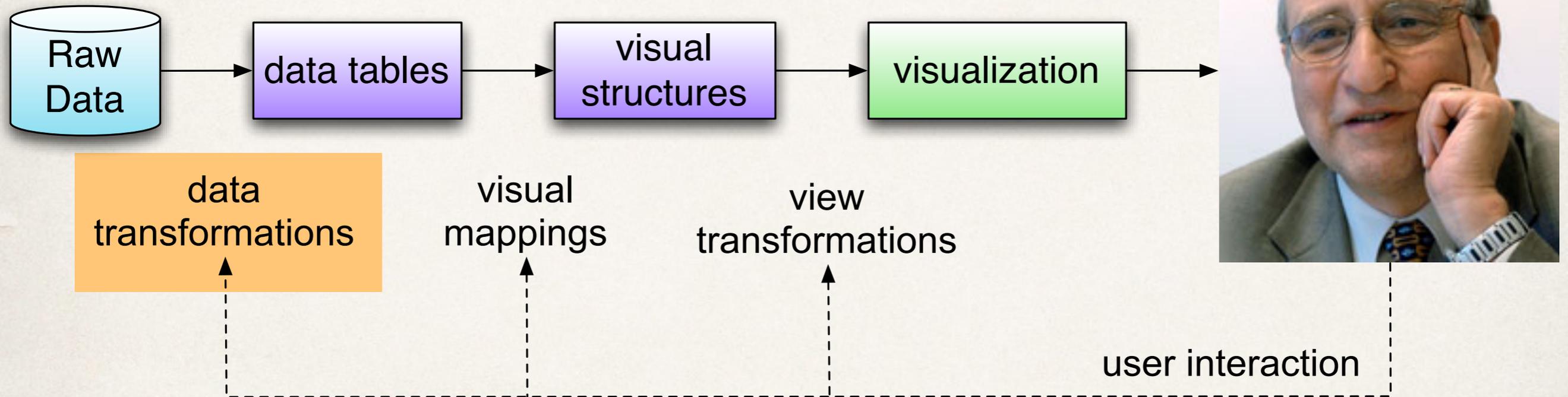
8 August 2008 – An unknown man visited the Lyari Town house where Bukhari is believed to stay. He arrived at 1615, and was let into the house immediately. Loud voices could be heard for a few moments, then they subsided. About fifteen minutes later a silver Mercedes left the rear of the house with what appeared to be three occupants. Due to the tinted glass it was impossible to identify the occupants of the vehicle. The house was surveilled for the next five hours; however no one came or went.

16 September 2008 – Bukhari was followed from his Lyari Town house to an apartment about 1.5 kilometers southeast. He entered the building and stayed there for about two hours. When he left he returned to his Lyari Town house at which time the observer believed he'd been discovered and left the area.

23 September 2008 – Bukhari and probably Basra are reported to have visited a house in the Katchi Abadis Old Settlement, on 835T Longhi Street. The informant, a vendor with business in the area, was passing by and noticed Bukhari and one other person enter the building at about 1430. The informant knew Bukhari by sight, but had never met him.

Visualization Pipeline

Insight!



Why is text visualization hard?

Text is high dimensional

(10,000+ unique words in a document)

Some words have ordering

January, February, March...

one, two, three...

Words have meaning,
relationships and context

On St. Patrick's Day, the local bar serves green beer.

The pub downtown has been known to make their pints chartreuse on Saint Paddy's .

Patrick walked around the corner of Green St. and Day St., right into this bar sticking out of the wall and knocked himself senseless.

Extracting structure

metadata - data about the data

who wrote it, when, what collection is it from, how long is it, format, keywords

lexical level - break the document into *tokens*

characters, words, n-grams

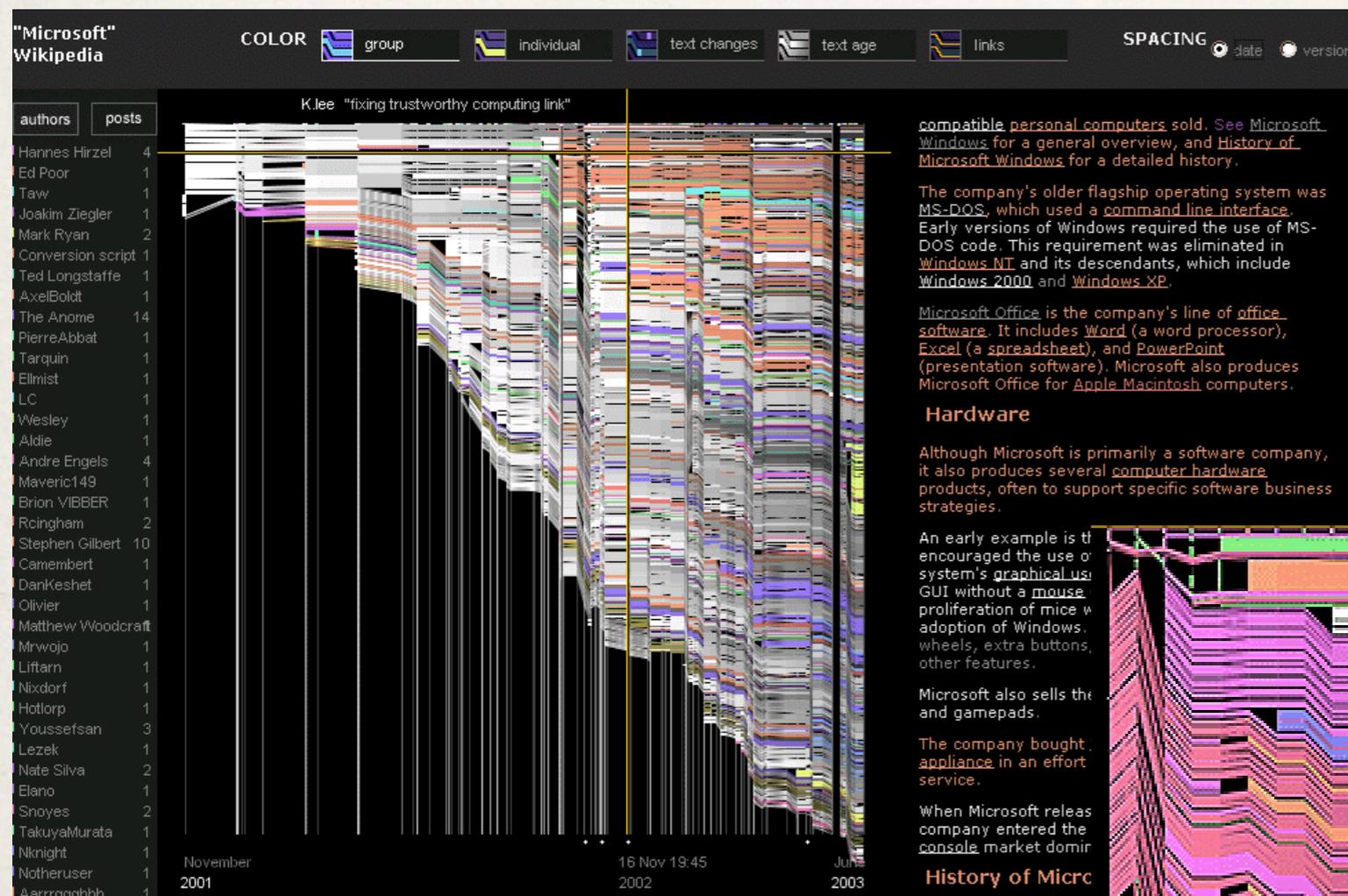
syntactic level - identify the function of tokens

identify the part of speech or named entities (people, places, money, dates)

semantic level - extract meaning from the text

identify relationships between entities, themes, sentiment, etc..

History flow



compatible personal computers sold. See [Microsoft Windows](#) for a general overview, and [History of Microsoft Windows](#) for a detailed history.

The company's older flagship operating system was MS-DOS, which used a [command line interface](#). Early versions of Windows required the use of MS-DOS code. This requirement was eliminated in [Windows NT](#) and its descendants, which include [Windows 2000](#) and [Windows XP](#).

[Microsoft Office](#) is the company's line of [office software](#). It includes [Word](#) (a word processor), [Excel](#) (a spreadsheet), and [PowerPoint](#) (presentation software). Microsoft also produces Microsoft Office for [Apple Macintosh](#) computers.

Hardware

Although Microsoft is primarily a software company, it also produces several [computer hardware](#) products, often to support specific software business strategies.

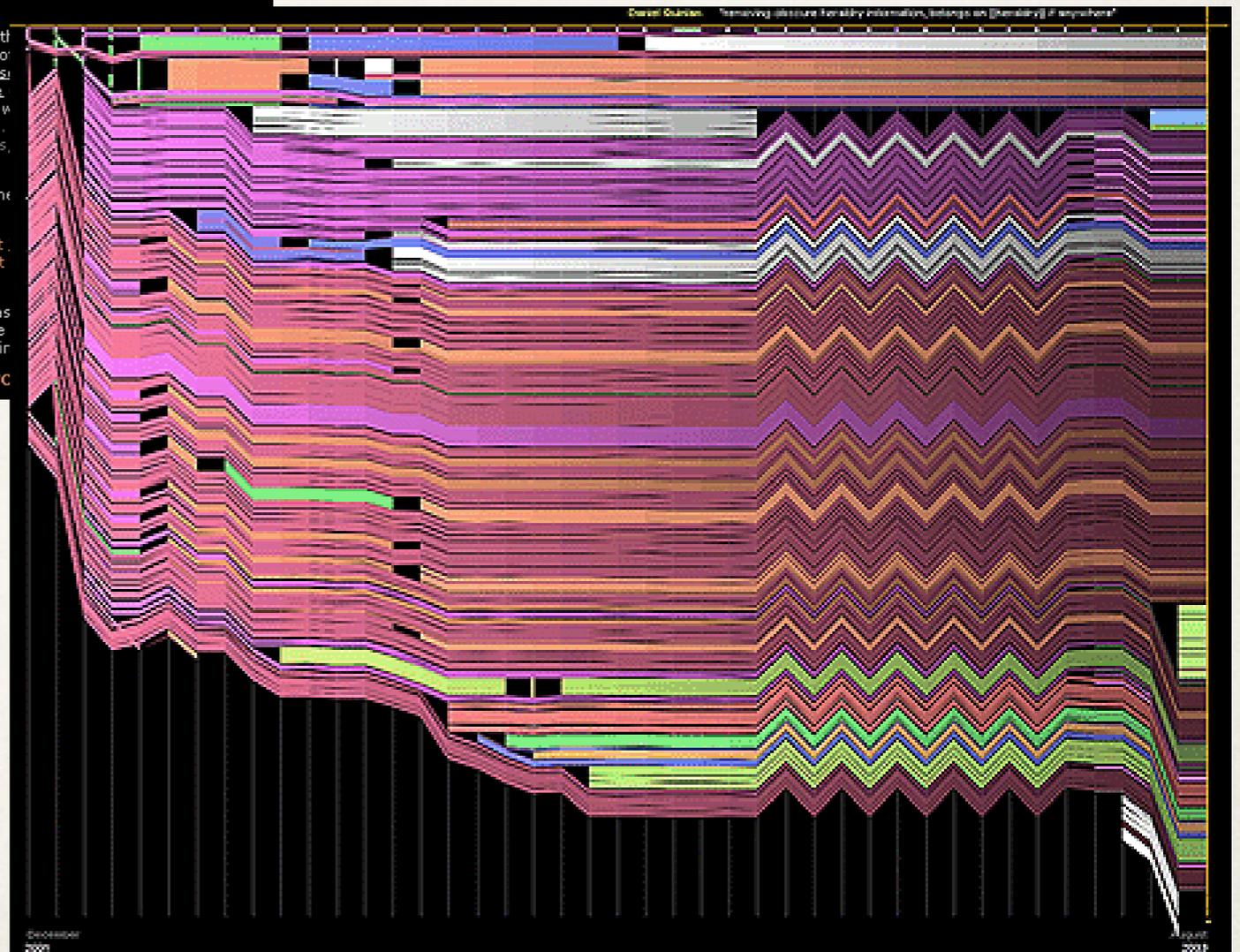
An early example is that encouraged the use of system's graphical user interface without a mouse. proliferation of mice w adoption of Windows. wheels, extra buttons, other features.

Microsoft also sells the and gamepads.

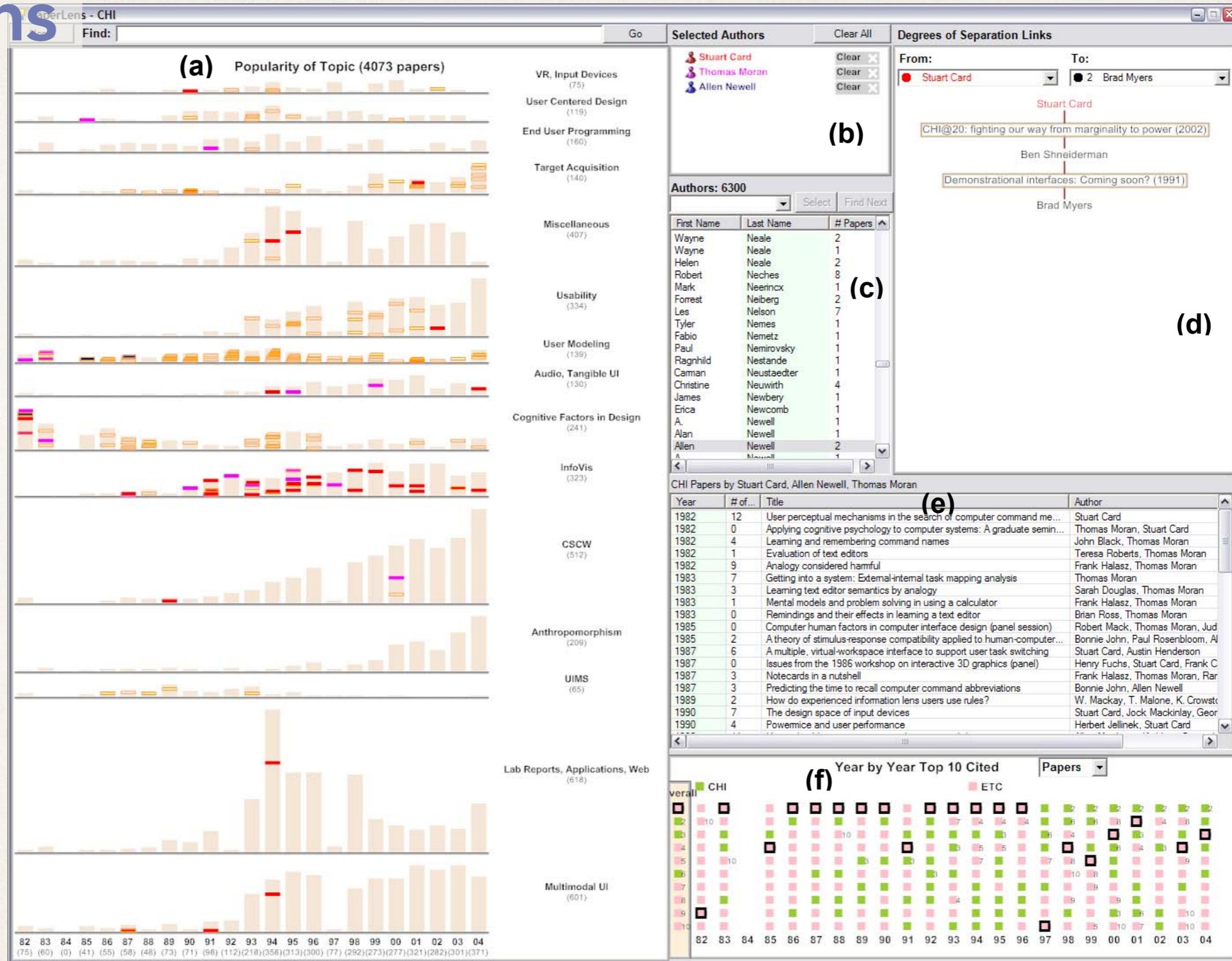
The company bought appliance in an effort service.

When Microsoft releas company entered the console market domin

History of Micro



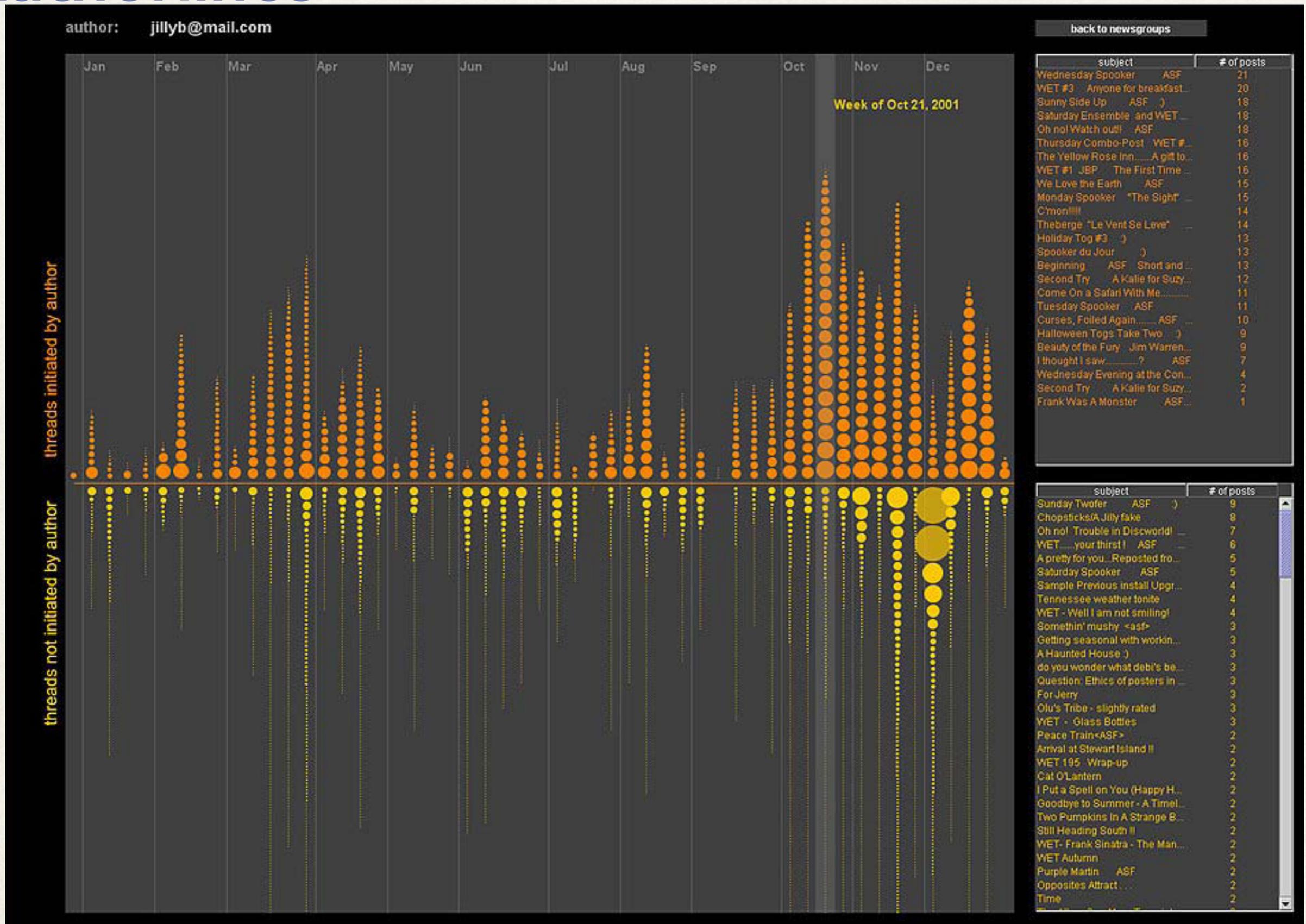
PaperLens



Lee et al., "Understanding Research Trends in Conferences using PaperLens"

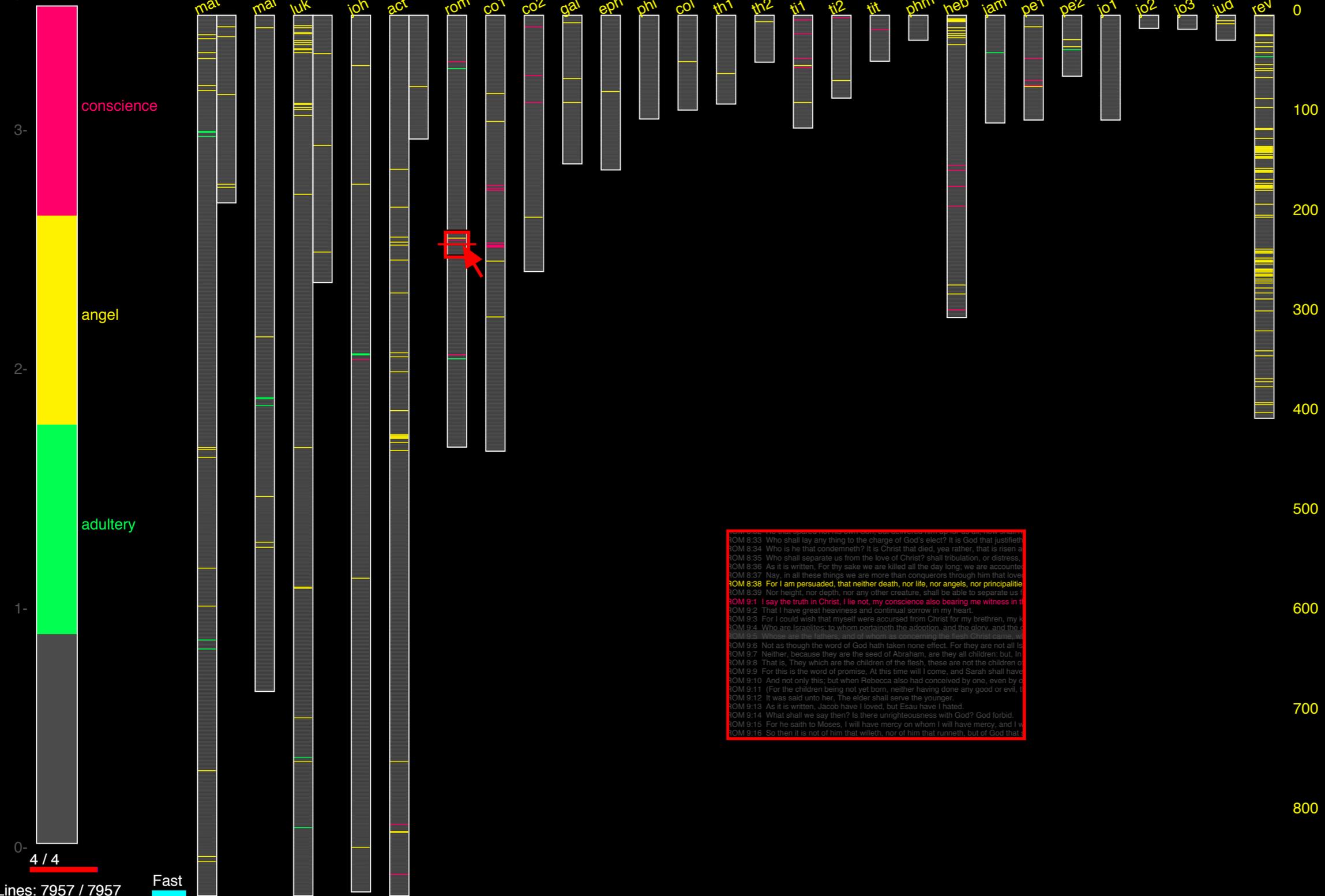
Figure 1. PaperLens tightly couples views across papers, authors, and references and consists of 6 main parts: (a) Popularity of Topic (b) Selected Authors (c) Author List (d) Degrees of Separation Links (e) Paper List (f) Year by Year Top 10 Cited Papers/Authors.

Authorlines



Viégas and Smith, "Newsgroup Crowds and Authorlines: Visualizing the Activity of Individuals in Conversational Cyberspace"

/tmp/words22058



ROM 9:5 Whose are the fathers, and of whom as concerning the flesh Christ came, we bless God for ever. Amen.

text: ROM 9:5 Whose are the fathers, and of whom as concerning the flesh Christ came, who is over all, God blessed for ever. Amen.

/tmp/words22058:

/tmp/words220

Vector Space Model / Bag of Words

	Book 1	Book 2	Book 3	Book 4
sir	481	79	118	12
about	316	175	192	61
jeeves	313	0	0	0
like	255	96	139	80
holmes	2	459	0	0
your	210	406	282	92
man	126	288	50	80
could	166	286	185	105
doctor	4	34	554	0
romana	0	0	419	0
k9	0	0	347	0
page	3	9	327	0
wizard	1	0	1	255
dorothy	0	0	0	238
jim	0	0	0	152
little	86	269	50	146
...

Weighting the vector-space

Term Frequency

$$TF_{td}(word) = count(word) \text{ in } d$$

Term Frequency by Inverse Document Frequency

$$TF.IDF_{td}(word) = TF_{td}(word) * \log \frac{N}{Df(word)}$$

Df() is # of documents containing the word, N is the number of documents

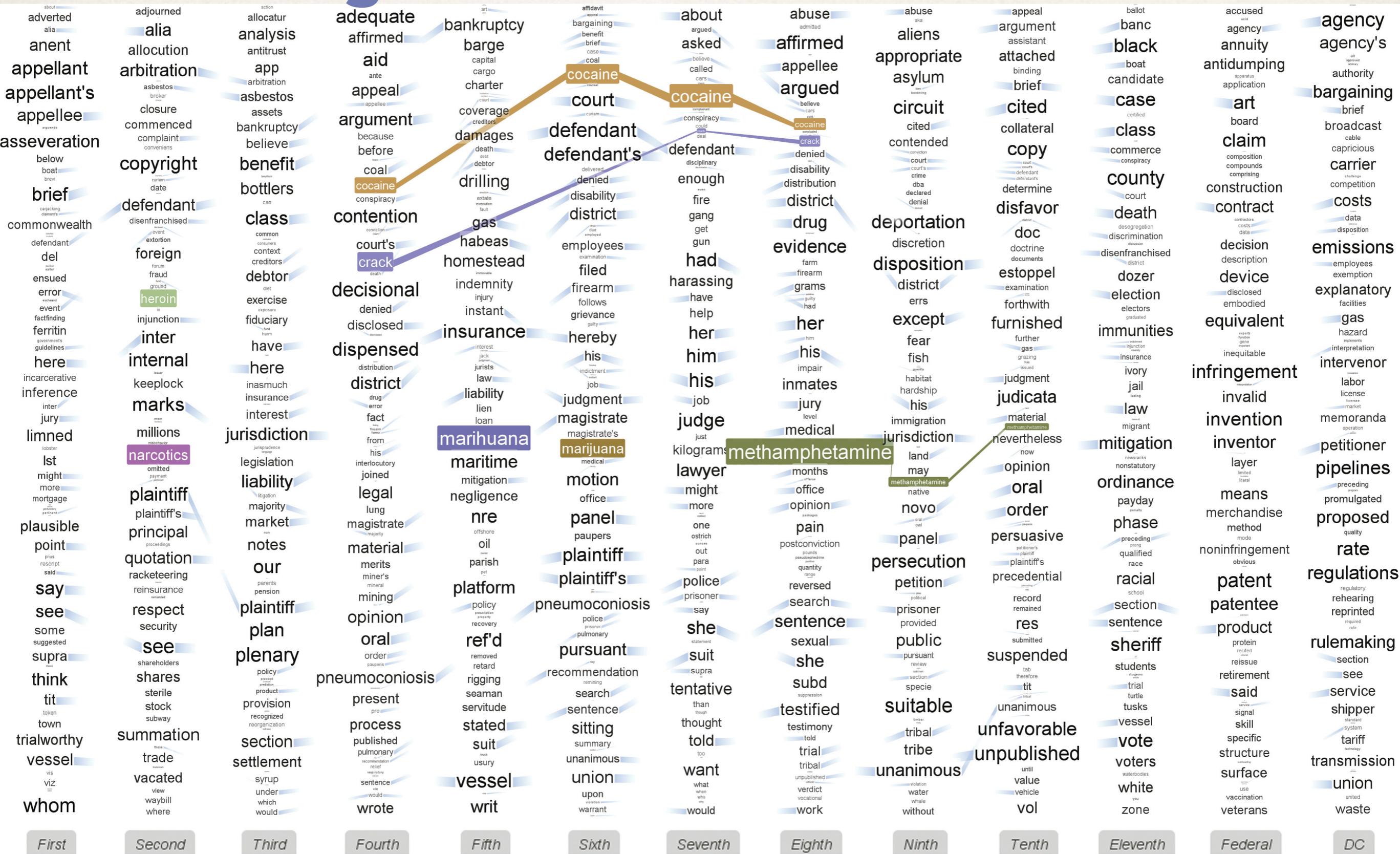
Tag cloud

_ absolutely afternoon anatole **angela** arrived asked ass augustus **aunt** back bassett bed bell **bertie** bertram **bit**
bloke bring **brinkley** business call **cannes** case **chap** child cold coming conversation country couple court cousin **dahlia** dash dashed day de dear
deeply **dinner** door **doubt** drones eh end engaged evening expect eye **eyes** face **fact** facts fear feel feeling feet fellow felt find fine
fink-nottle fire form **found** friend garden gave **girl** give glossop **good** great **gussie** hadn half hand hands
happened happy hard **head** hear **heard** heart ho home hour house hullo human idea imagine **jeeves** juice kind knew left **life** long
looked lost **love** **made** make making **man** manner market **matter** men met **mind** mine minutes miss **moment** morning **mr**
newt newts **night** note occasion open opinion order part passed people pie **place** point poor position possibly pretty prizes **put** read remember
rest room **round** scarcely scheme **school** set shark **side** sight simple simply **sir** snodsbury **sort** soul speak speech spoke spot stand standing
start started stood story stuff suddenly **suppose** taking talk talking **thing** things thinking **thought** till **time** told tom
tonight touch travers true **tuppy** turned **uncle** understand voice woman **wooster** word words work world wrong years young

Showing 200 out of 7355 words

Right ho, Jeeves

Parallel Tag Clouds



Collins, "Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora"

Word Tree

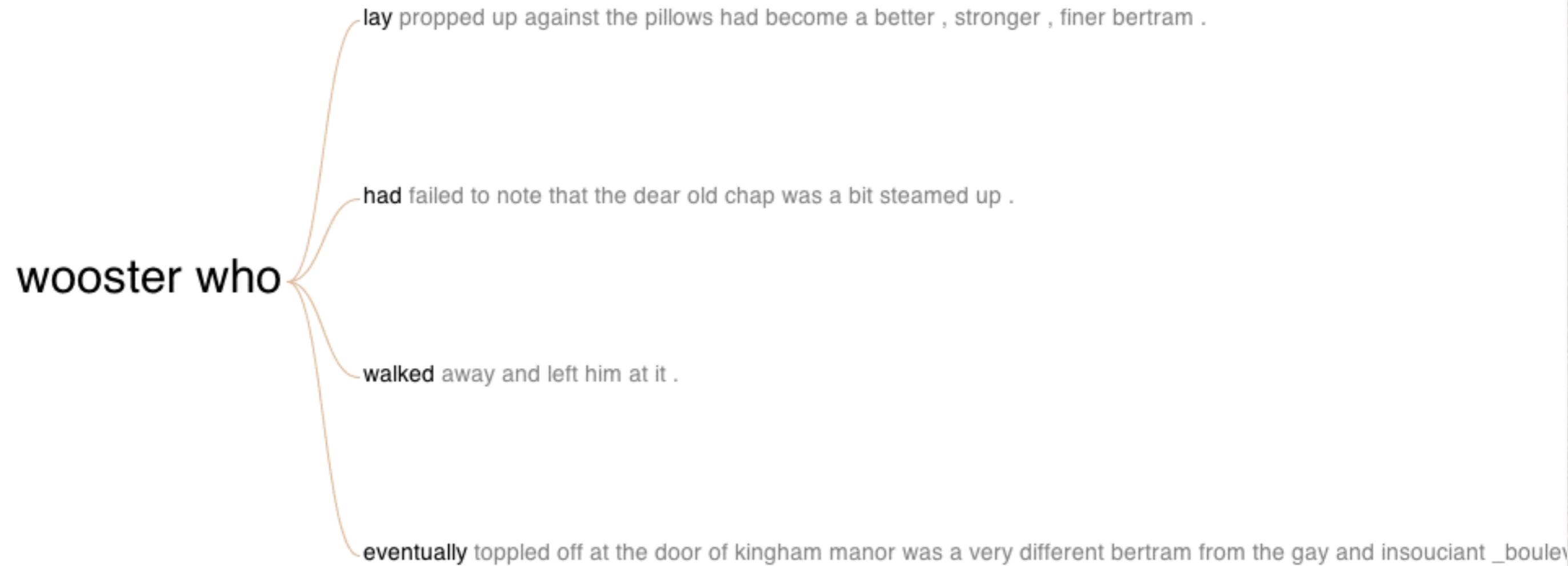
51
hits

wooster



Word Tree

4
hits



Phrase net

Showing 36 of 650 terms

Select a phrase

- word1 and word2
- word1 's word2
- word1 of the word2
- word1 the word2
- word1 a word2
- word1 at word2
- word1 is word2
- word1 [space] word2

or enter your own

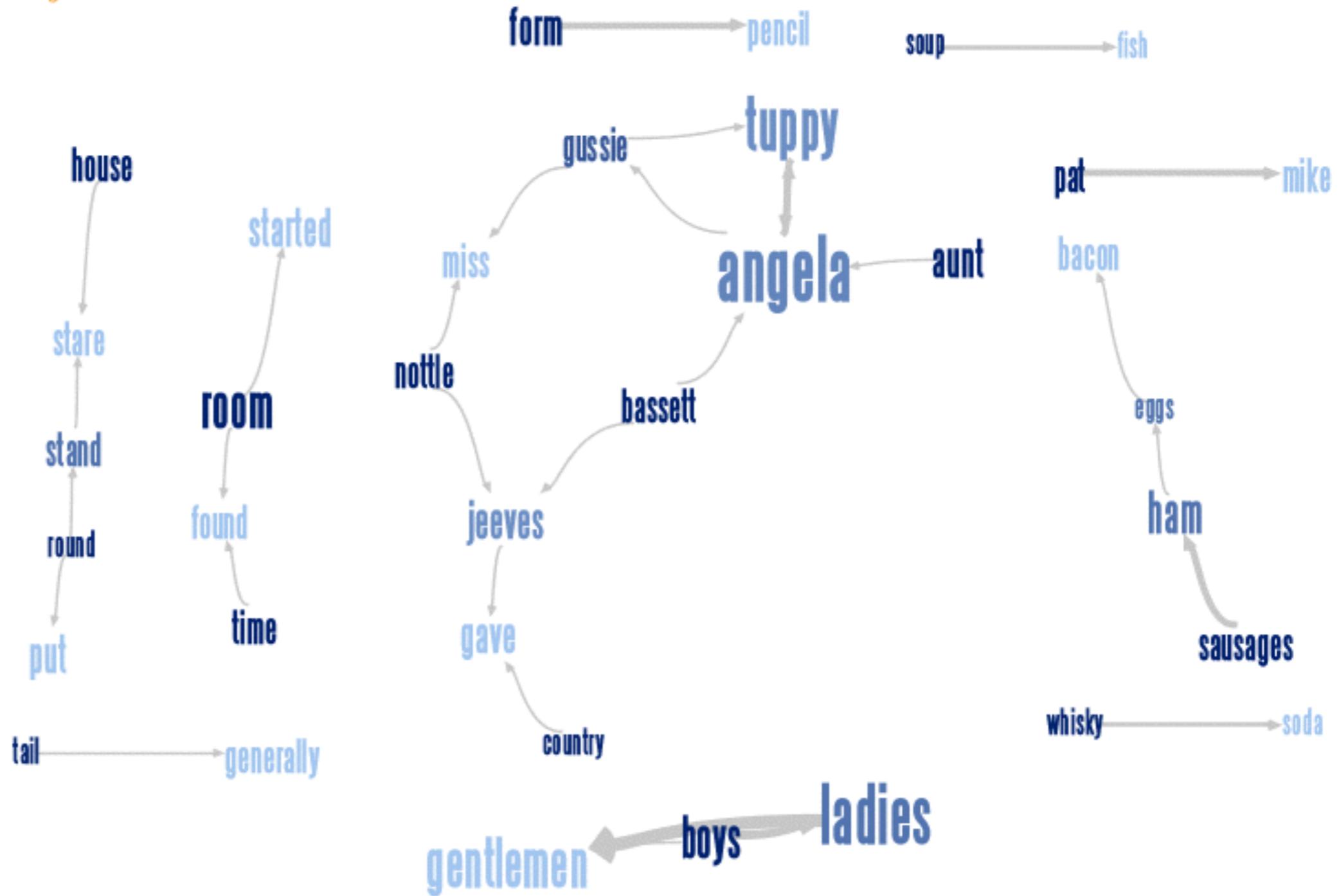
Filters

Show top:

Hide common words

Zoom

In Out Reset



FeatureLens

Load About
FeatureLens

Frequent Patterns

Filtering

Pattern contains :

Search patterns

Order patterns by

Frequency :

Length :

Trends :

Trends per section :

Append pattern to legend

- // tonight
- 77 americans
- 75 security
- 74 congress
- 68 government
- 65 make
- 64 years
- 62 work
- ✓60 freedom
- 57 great
- 56 united
- 54 good
- 53 citizens
- 53 children
- 52 time
- ✓52 terrorists
- 51 states
- 50 economy
- 50 terror
- ✓48 war
- 44 iraq
- 43 the united states

previous 100 next 100

Load history go

Collection Overview : 'The State of the Union' (1 doc per line - 140 hiligh

Sections Overview reset

2001-1

2001-2

2002

2003

2004

2005

2006

2007

Legend

- he has not|that sadd + X

- freedom + X

- terrorists + X

- war + X

Document View

Show Selection and context Results only

55

The United Nations concluded in 1999 **that Saddam Hussein had** biological weapons materials sufficient to produce over 25,000 liters of anthrax - enough doses to kill several million people. **He has not accounted for** that material. **He has given no evidence that he has**

56

The United Nations concluded **that Saddam Hussein had** materials sufficient to produce more than 38,000 liters of botulinum toxin - enough to subject millions of people to death by respiratory failure. **He has not accounted for** that material. **He has given no evidence that he has destroyed it.**

57

Our intelligence officials estimate **that Saddam Hussein had** the materials to produce as much as 500 tons of sarin, mustard, and VX nerve agent. In such quantities, these chemical agents also could kill untold thousands. **He has not accounted for** these materials. **He has given no evidence that he has destroyed** them.

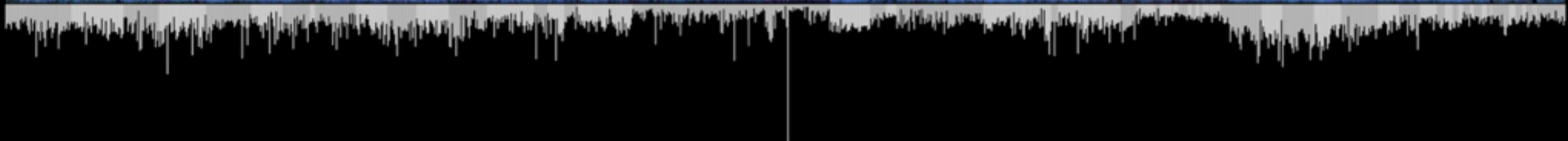
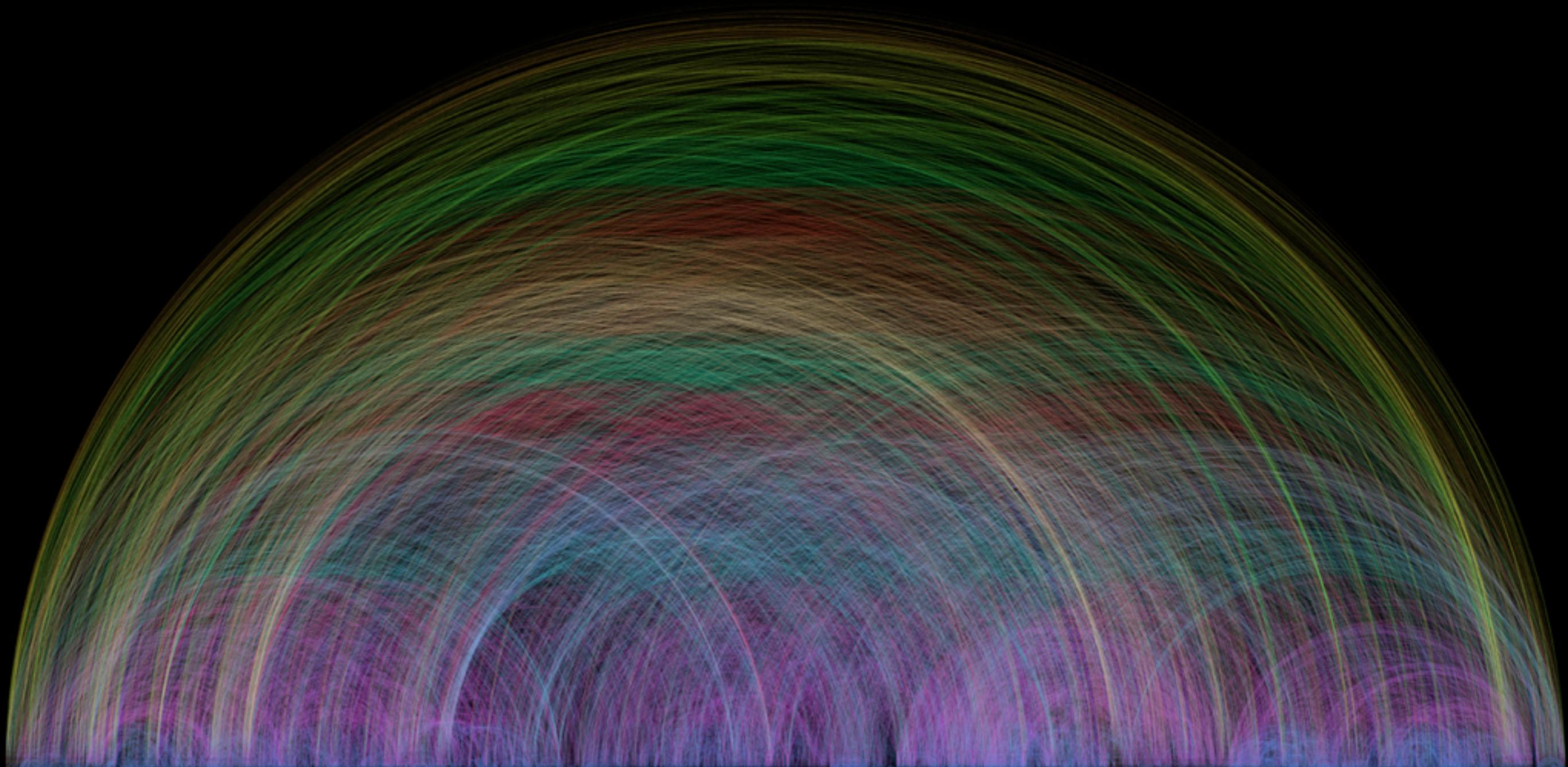
Don et al.,
 "Discovering interesting usage patterns in text collections: integrating text mining with visualization"

Google nGram Viewer

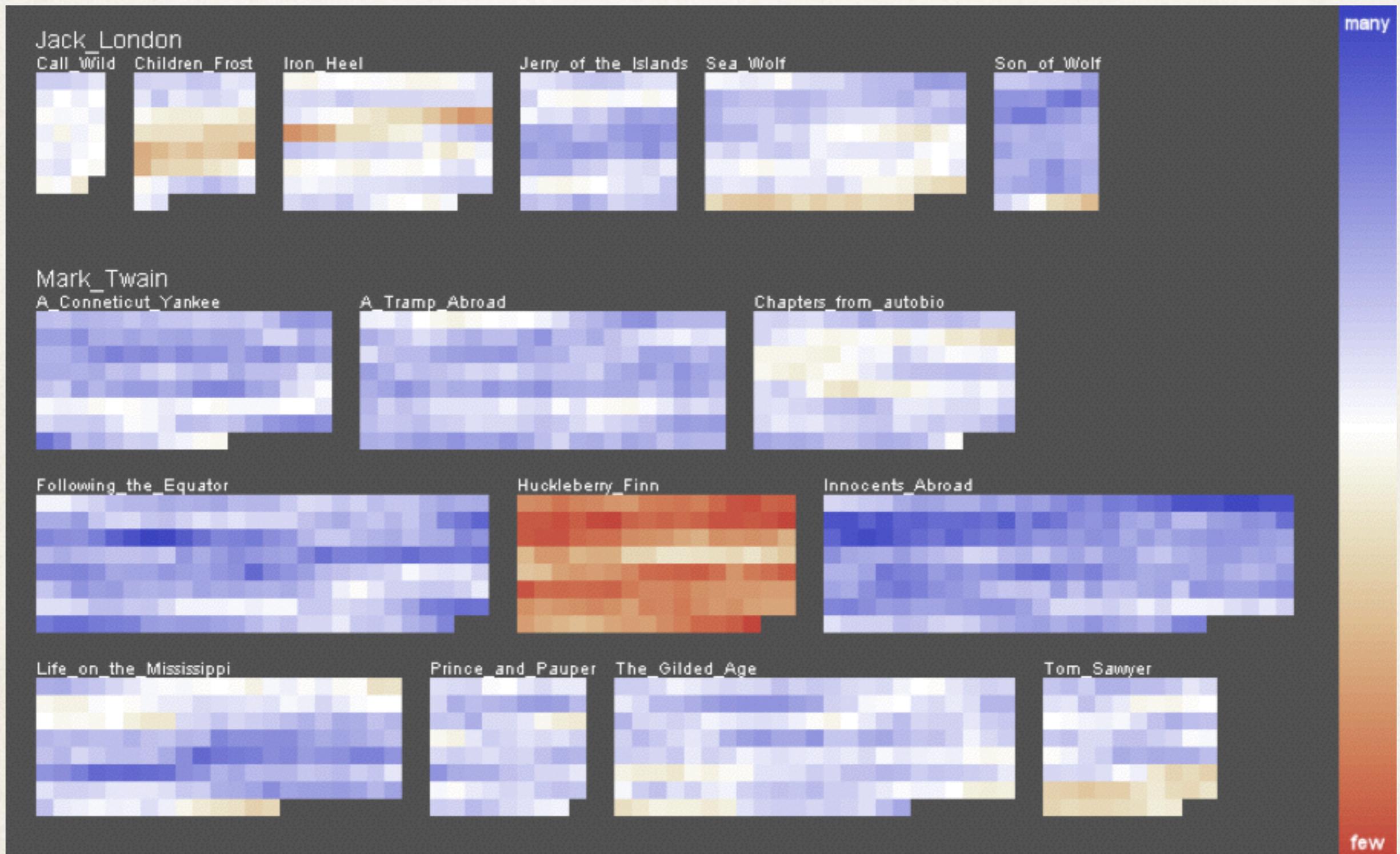
Graph these comma-separated phrases: case-insensitive
between and from the corpus with smoothing of .



(click on line/label for focus)



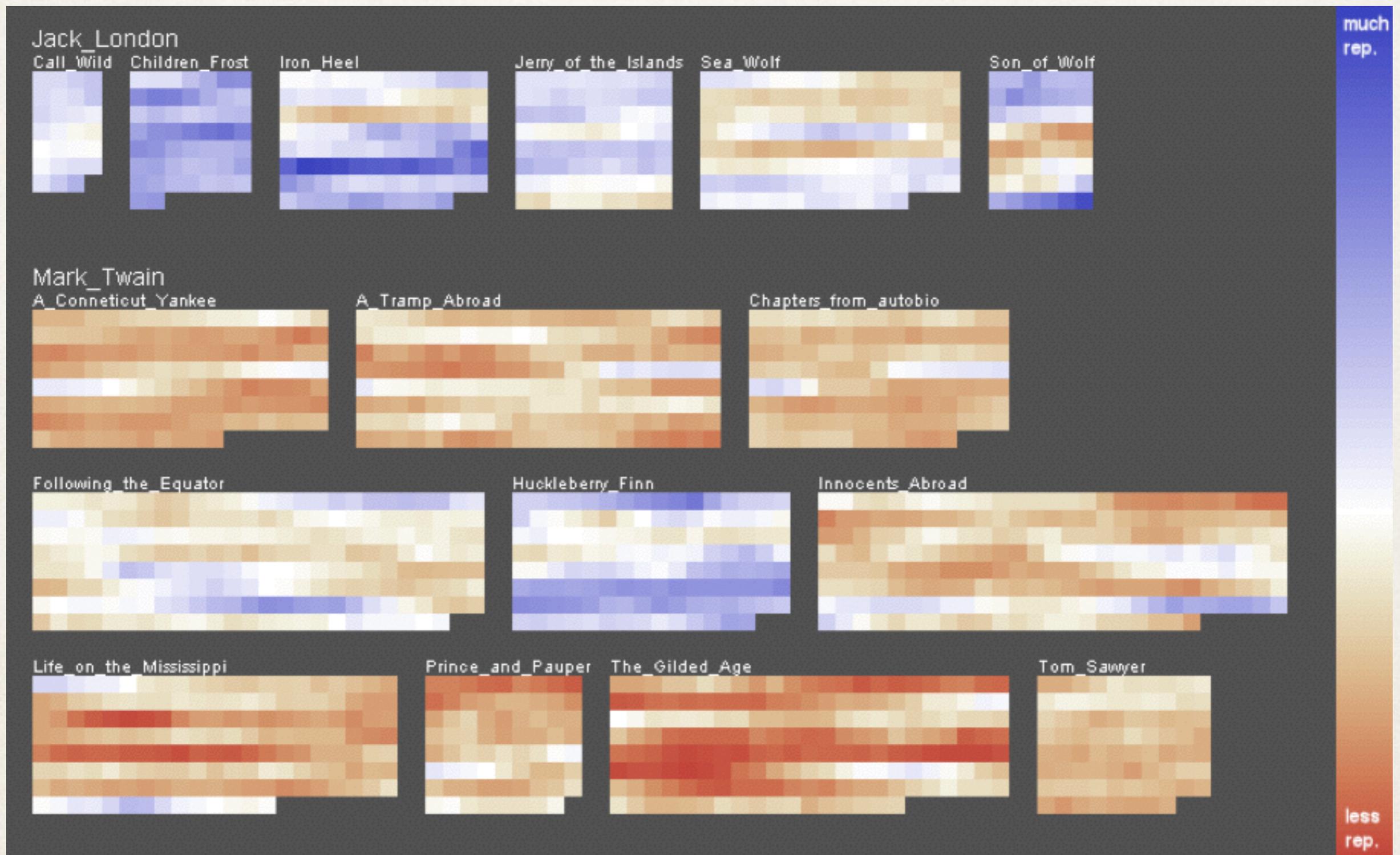
Literature fingerprinting



Hapax Legomena
word that appears only once

Keim, Oelke, "Literature Fingerprinting: A New Method for Visual Literary Analysis"

Literature fingerprinting

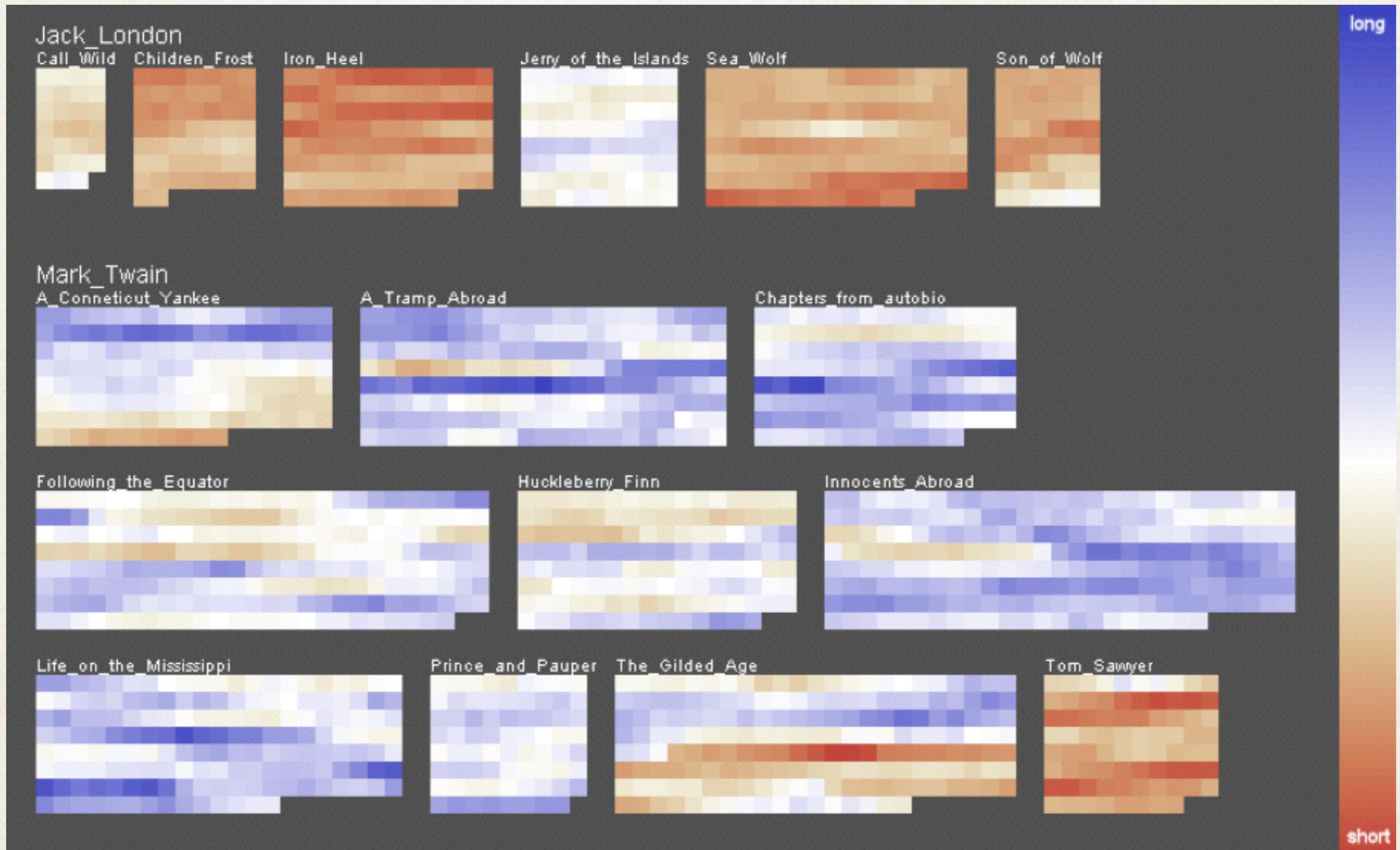


Simpson's Index

probability that a token belongs to the set

Keim, Oelke, "Literature Fingerprinting: A New Method for Visual Literary Analysis"

Literature fingerprinting



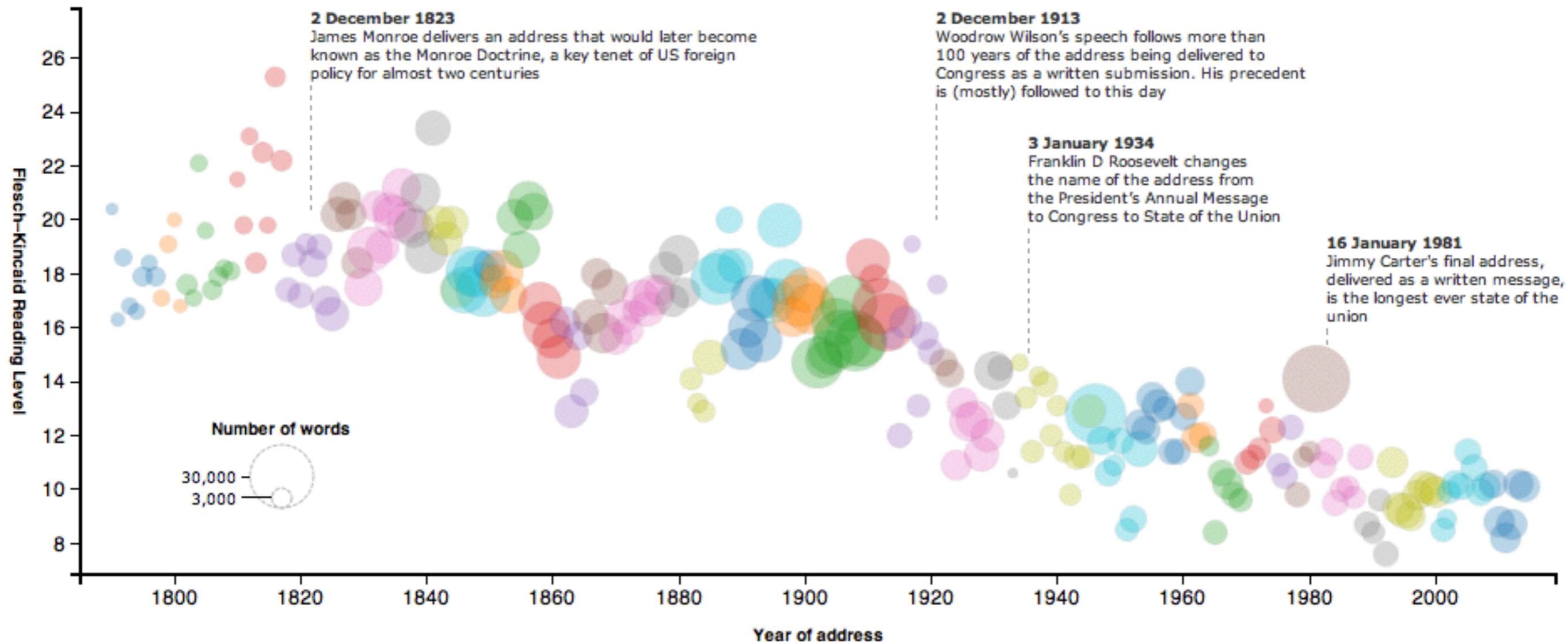
Average sentence length

Keim, Oelke, "Literature Fingerprinting: A New Method for Visual Literary Analysis"

Reading level

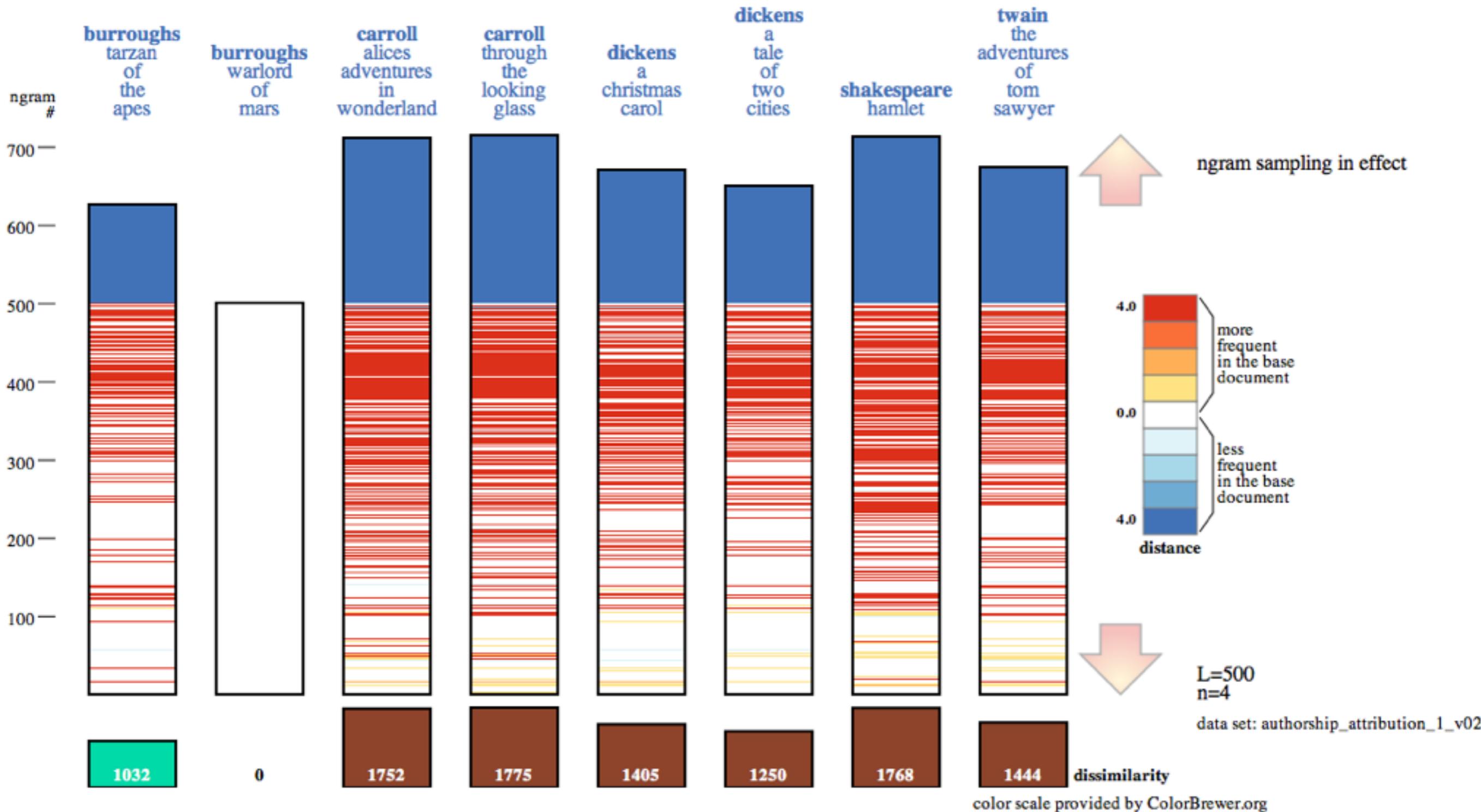
The state of our union is ... dumber: How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every State of the Union

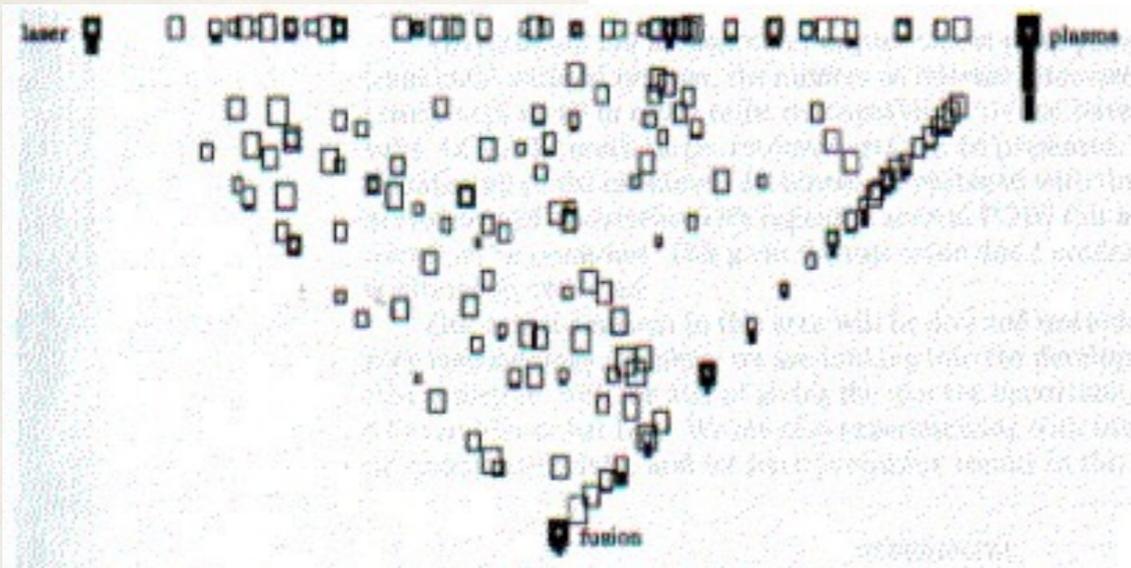
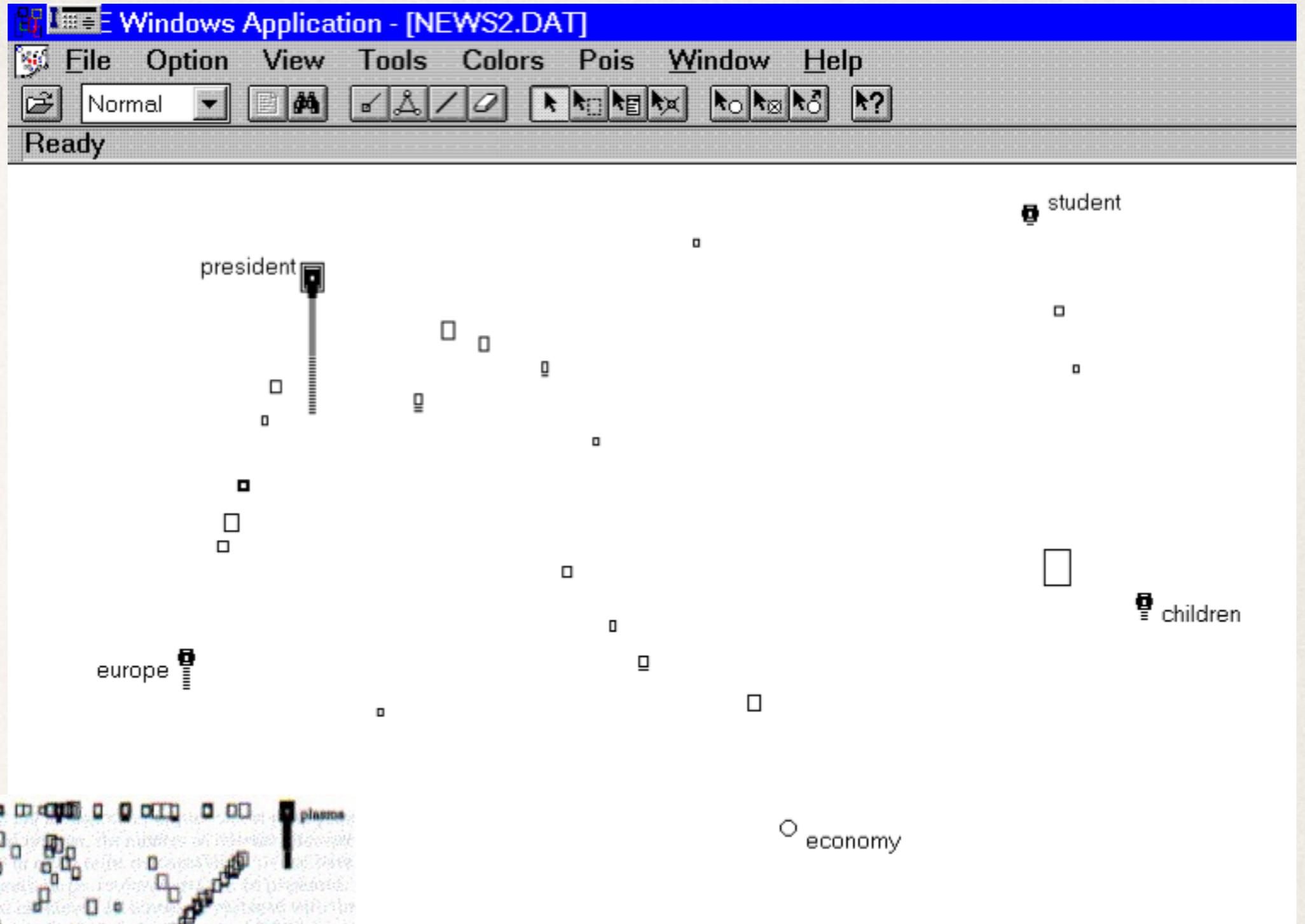


Relative n-gram signatures

base document: **burroughs warlord of mars**



VIBE



Review Spotlight

'09 amazing around baked bar bass best chef delicious eat
elite everything favorite fish food fresh going hamachi
hawaiian hour line love mango minutes mussels name
night nigiri order people prices really restaurant roll
expensive or cheap? sake salmon sea seated service spicy
table think tuna wait waitress W

“long wait” or “no wait”?

b) best sf
baked sea bass best sushi sure in striped bass
other person
fresh fish slow service sushi bar
sushi chef baked mussel more hour only thing
long wait long time sushi restaurant good food
long line hawaiian roll reasonable price
baked mango small place delicious everything

Mentioned 63 times

possess sage of the halos wisdom , and know in advance sushi zone only accepts cash and the waits will be long and arduous .

yes , its a long wait , learn the master of zen if you want to eat here .

we came here early to try to avoid the long wait most people here talk about .

Named Entity Extraction

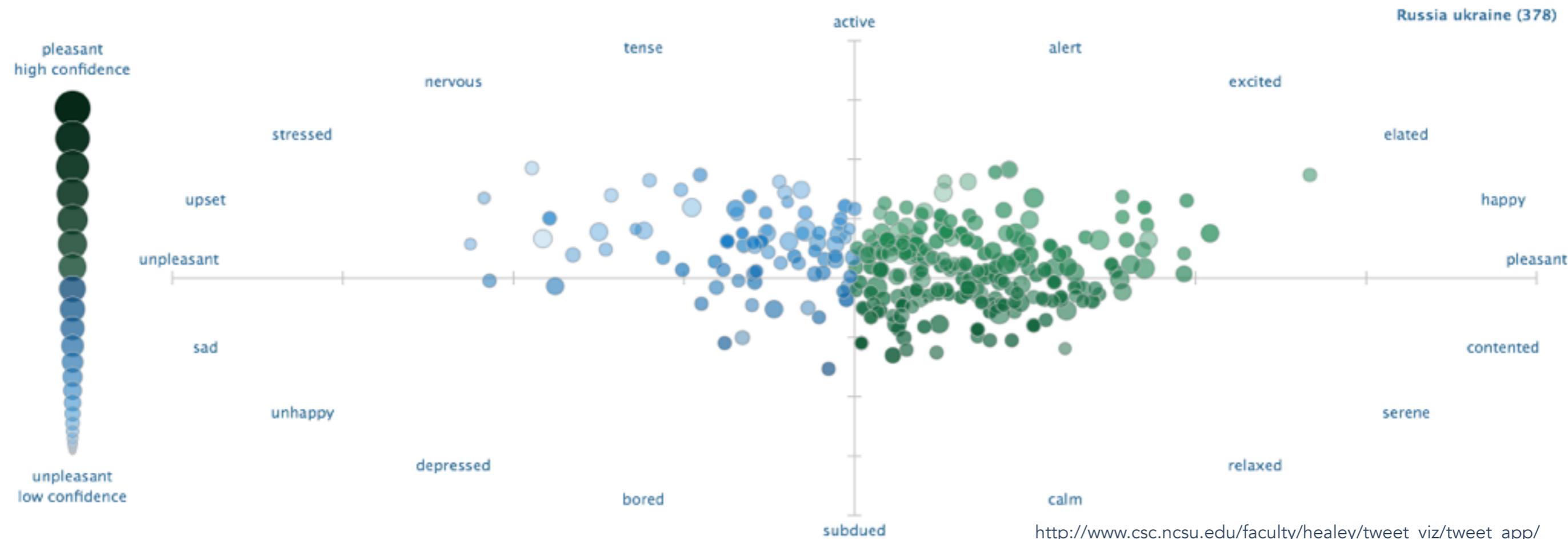
A senior police officer was killed near **Cairo** on **Wednesday** in a bomb attack claimed by a militant group, while security forces stormed a hideout used by another Islamist organisation near **Alexandria** in a raid that left an officer and a militant dead. Militant violence has spiralled since **last July**, when the army toppled elected head of state **Mohamed Mursi** and the authorities launched a fierce crackdown on his supporters in the Muslim Brotherhood and other Islamist sympathisers. The attacks underline lingering instability in **Egypt** ahead of a presidential election in **May** that **Abdel Fattah** al-Sisi, the former army chief who deposed **Mursi**, is expected to win. The prime minister said the state was in "a fierce war" on terror. The police officer killed near **Cairo** was named as Brigadier General **Ahmed Zaki**. State media said he was killed outside his home in 6th of **October** City, 32 km (20 miles) outside **Cairo**, when a bomb placed under his car went off. Two conscript policemen were wounded in the bombing. A militant group called **Ajnad Misr**, or Soldiers of **Egypt**, said it carried out the attack in a statement posted on a Facebook account in its name that has carried past statements. The post included a photo of a man said to be **Zaki** on his way to his vehicle, describing him as "the criminal brigadier general in the (security) force for killing protesters". Mursi's removal from power **last summer** after mass protests against his rule tipped **Egypt** into the worst internal strife of its modern history. Hundreds of his supporters were killed by security forces as they broke up their protest camps. Militant attacks since then have killed around 500 people, mostly policemen and soldiers. The threat has been compounded by a flow of weapons from neighbouring **Libya**. The **Interior Ministry** said the hideout targeted by police at dawn on **Wednesday** near **Alexandria** was used by **Ansar Bayt al-Maqdis**, or Supporters of **Jerusalem**, the group behind some of the deadliest attacks of the last nine months. The militants had opened fire on the security forces as they arrived at the hideout in **Borg El Arab**, some 45 km (28 miles) south-west of **Alexandria**. The police officer killed in the raid was named as **First Lieutenant Ahmed Saad** and the dead militant as **Hassan Abdel Aal**, a 25-year old from the **Nile Delta** province of **Dakahlia**. Two other militants were detained, the ministry spokesman, **Hany Abdel Latif**, said in a televised statement. Footage broadcast on state TV appeared to show the body of a militant on the ground. The militants were "among the dangerous elements of the terrorist group **Ansar Bayt al-Maqdis**, which was planning to target police and military facilities and the security forces", the ministry said. The police seized weapons including explosive belts, automatic weapons, hand grenades and ammunition.

Potential tags:

LOCATION
TIME
PERSON
ORGANIZATION
MONEY
PERCENT
DATE

<http://uk.reuters.com/article/2014/04/23/uk-egypt-violence-idUKBREA3M1KO20140423>

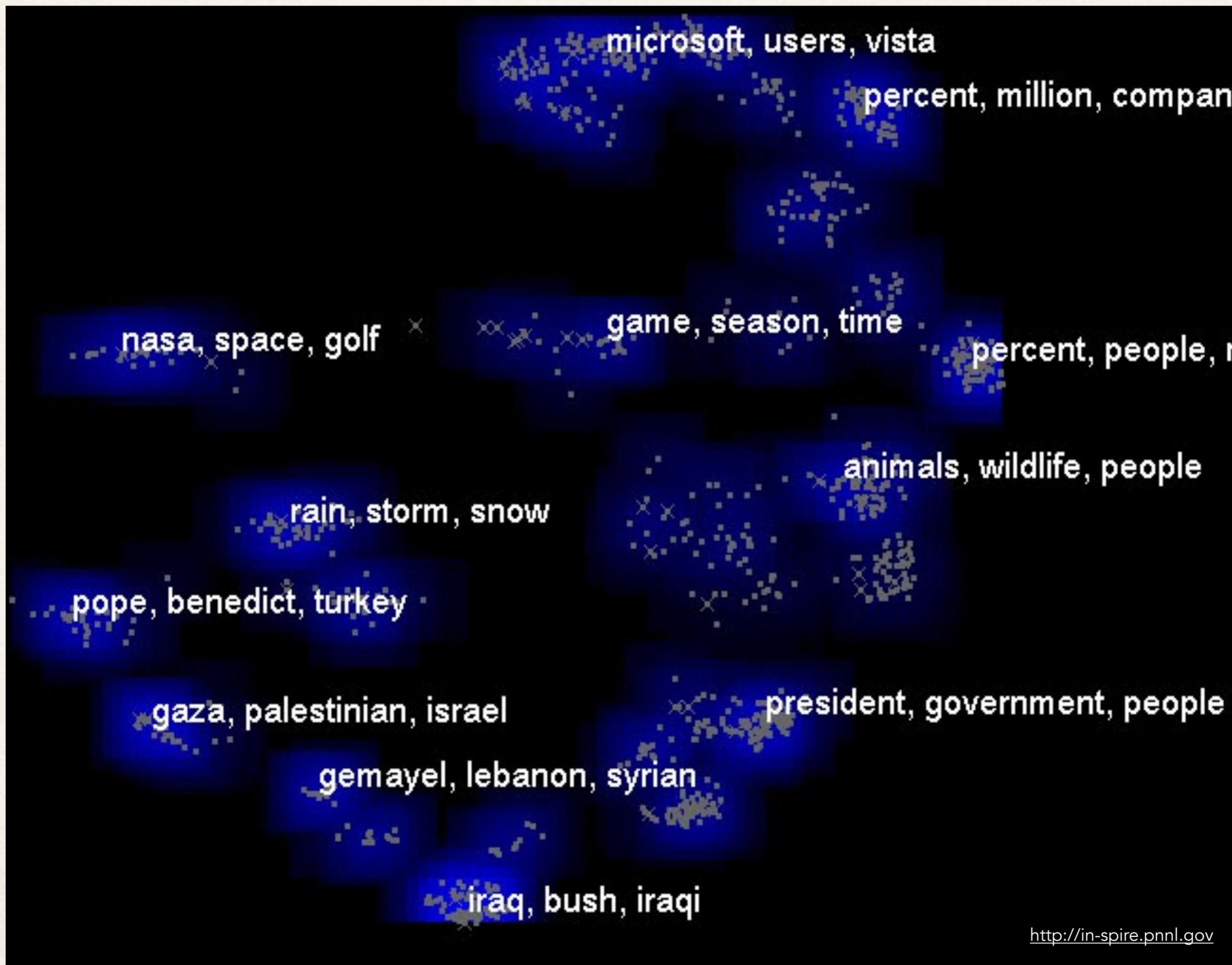
Sentiment analysis



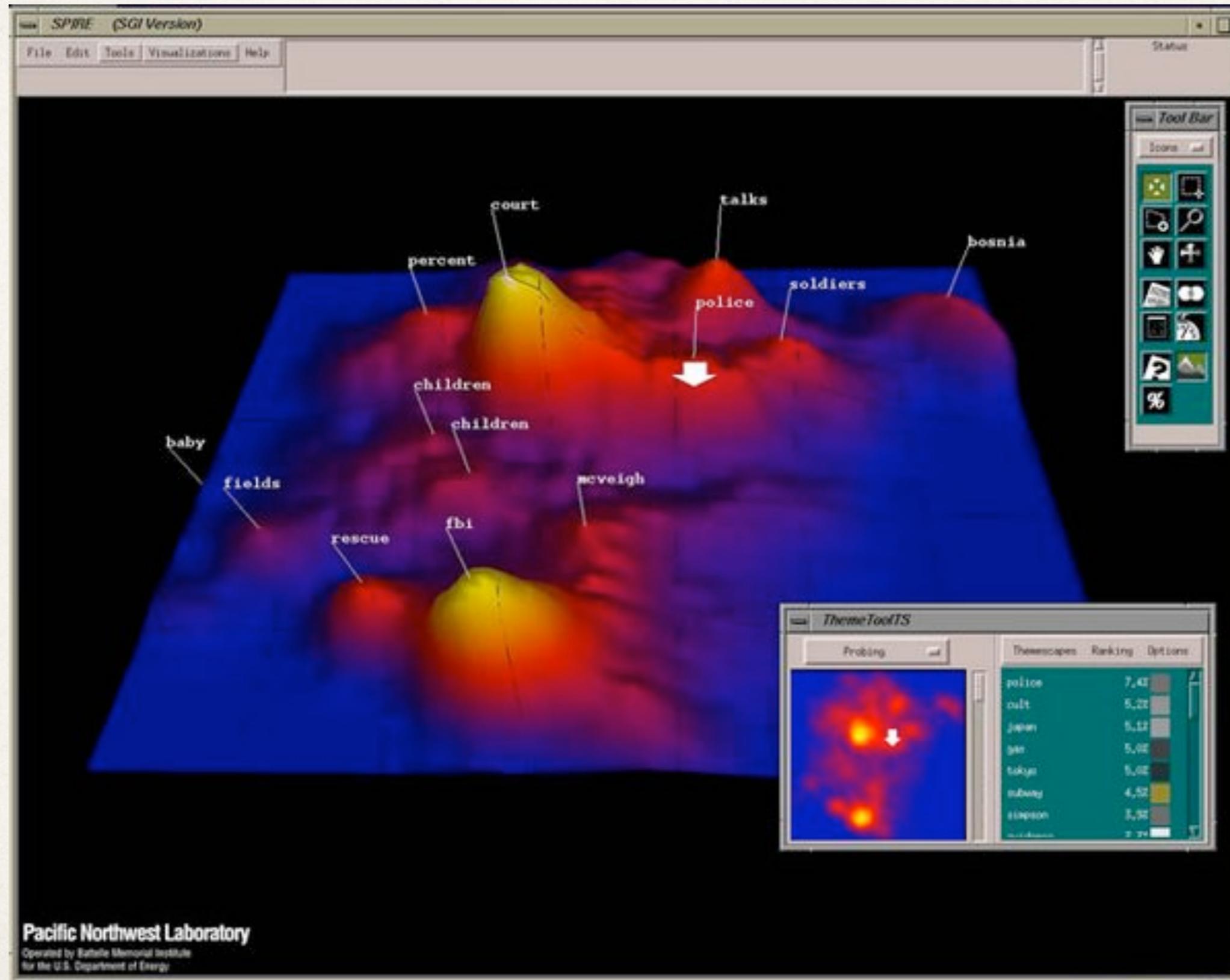
http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Date	User	v	a	Tweet
04-23-14 21:56	NaijaAviators	5.36	4.79	In an <i>apparent show of force</i> to #Russia, #U.S. troops have landed in Poland http://t.co/isDqsvkxM6 #Ukraine via @cnni @omojuwa
04-23-14 21:56	SecretNews	5.21	5.42	Sergey Lavrov: 'If Russian Troops Or <i>People Attacked</i> , We'll Retaliate': http://t.co/P2k4FCc9Nj #breakingnews,# <i>breaking</i> ,# <i>politics</i> ,#military
04-23-14 21:56	VNBreakingNews	5.19	5.64	Russia says it will <i>respond</i> if Ukraine <i>interests attacked</i> at http://t.co/F75YVJhkIc
04-23-14 21:56	jackdrsm	5.92	3.96	RT @GregRaths: Is Ukraine the <i>starting point</i> of the next Exodus? http://t.co/KILQlcDAJp #Ukraine #Russia #america #americanidol
04-23-14 21:58	YokelChang	6.34	4.61	#YokelChang Vladimir Putin's <i>shopping list</i> : Which country could be next?: <i>Plenty</i> of Russia's neighbors are wor... http://t.co/WjPq2ARDwE
04-23-14 21:58	AndresSuurkuusk	3.88	5.35	#Ukraine Resumes <i>Bid</i> to Oust Militants Amid #Russia'n # <i>Threat</i> http://t.co/1tyBIKRqU7 via @BloombergNews
04-23-14 21:58	MockCasbah	5.87	4.94	@miodragsoric @AtlanticCouncil @MKasyanov <i>Well</i> first we need to <i>make sure</i> Ukraine <i>makes</i> it that long. Russia is <i>mighty hungry</i> .
04-23-14 21:58	SonyaHeaney	5.96	4.76	RT @peterbakernyt: <i>Fascinating look</i> at <i>links</i> btwn <i>mysterious "green men"</i> in Ukraine and Russian troops. @AndrewKramerNYT Higgins Gordon htt...
04-23-14 21:59	Spiritof1642	6.15	5.55	Russian <i>Foreign</i> Minister Sergey Lavrov <i>gives</i> Russia Today interview, in English, re #Ukraine. http://t.co/qSqqbo8nYT
04-23-14 21:59	taniaLucyjade	5.89	5.23	RT @EuromaidanPR: #Russia Warns #Ukraine of Potential Military <i>Response</i> - <i>New York Times</i> http://t.co/x4iZ5C1vas @BlogsofWar EMPR <i>News</i>
04-23-14 21:59	GregoryHSmith2	5.39	5.11	"@cnni: In an <i>apparent show of force</i> to Russia, U.S. troops have landed in Poland: http://t.co/Cb9vWhRxyf " <i>Growing risk</i> of violence
04-23-14 21:59	euromaidantwit	5.28	5.57	Hillary Clinton <i>calls</i> for more sanctions on Russia& she <i>believes</i> the <i>outcome</i> in #Ukraine will be a <i>bad one</i> for Russia http://t.co/ecNLRwdE4e
04-23-14 21:59	LanellePenafior	5.04	3.89	Amid Russia <i>warning</i> , Ukraine is in a <i>security bind</i>
04-23-14 21:59	josephwosk	5.31	6.18	RT @Thoreau_H_D: Russia: Duma(<i>Congress</i>) to <i>recognize</i> the 'Independence' of Transnistria, bordering on Moldova/Ukraine. #euromaidan http://t.co/...
04-23-14 22:00	art2x2	6.12	5.12	As U.S. <i>Army</i> <i>Bad</i> <i>Boats</i> Entered Poland, They Were <i>Created</i> on the Airfield by <i>Polish</i> Soldiers #P2 #Ukraine #Russia http://t.co/...

In-Spire: Galaxy view



In-spire: Theme view [classic]



In-Spire: Theme view

