Introduction to text mining (with R and tidyverse)

Peeter Tinits (University of Tartu/Tallinn University)

03.04.2019 at Rīgas Tehniskā Universitāte

What is text mining

- (textual) data <-> questions

- Questions matter.
- Technology can be easy (if data is available)

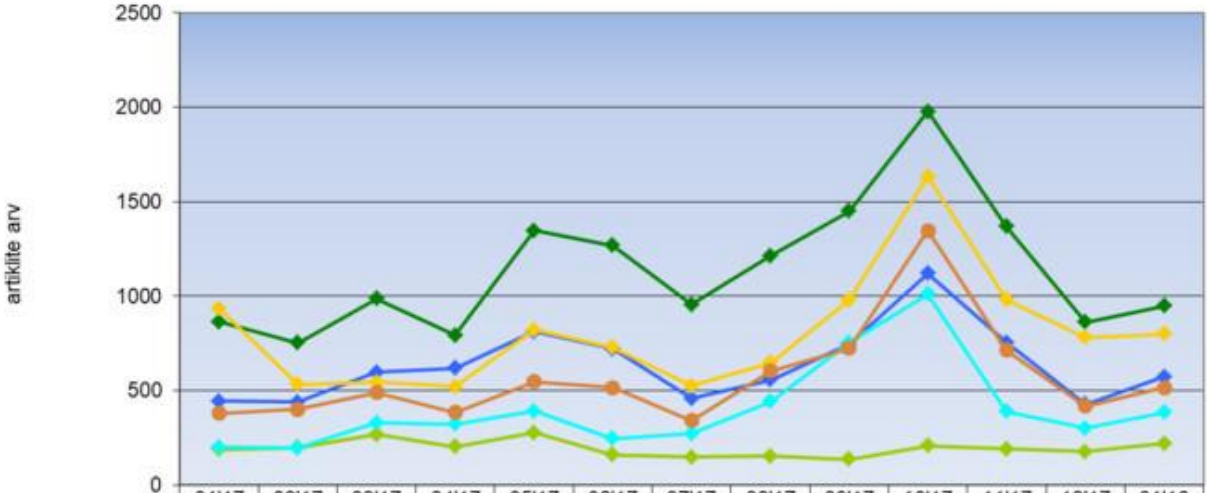
Examples of text mining

Media monitoring

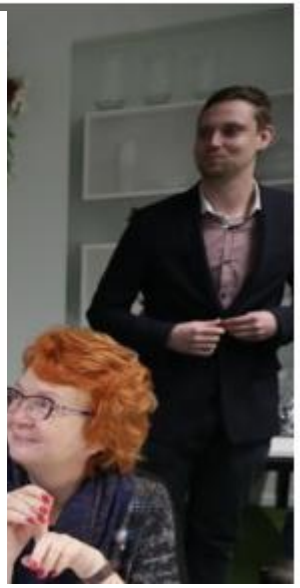
Jaanuaris pälvis meedias enim tähelepanu Keskerakond

EESTI

18.02.2018 16:51

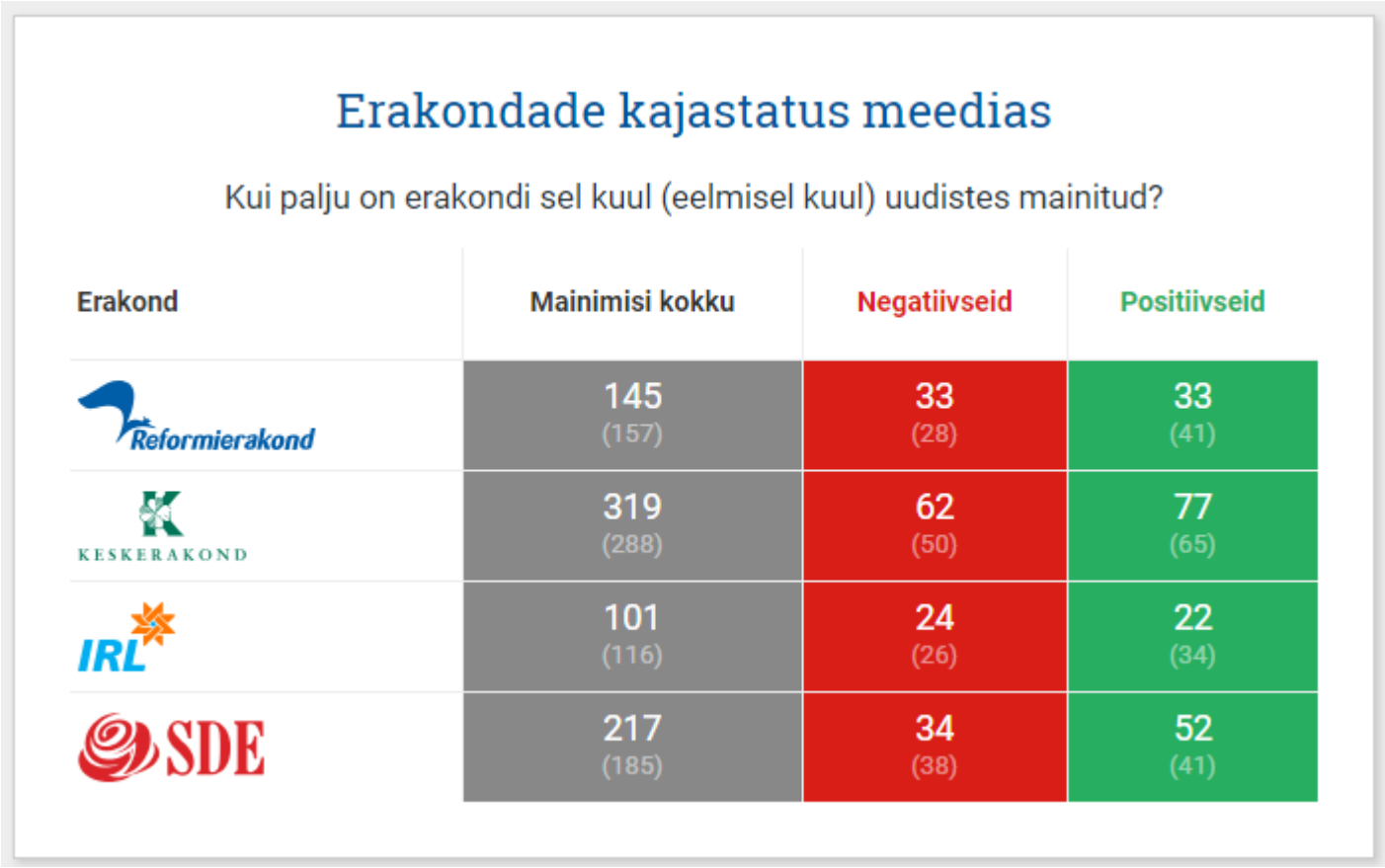


	01'17	02'17	03'17	04'17	05'17	06'17	07'17	08'17	09'17	10'17	11'17	12'17	01'18
Vabaerakond	188	197	267	202	278	161	147	152	137	207	190	177	220
IRL	443	440	598	618	811	721	459	557	746	1121	755	427	574
Keskerakond	864	753	987	793	1348	1267	954	1212	1450	1978	1368	863	947
EKRE	201	196	329	322	392	247	272	442	754	1010	390	302	385
Reformierakond	936	532	544	521	821	728	526	646	979	1634	982	780	799
SDE	380	401	489	385	547	516	342	603	724	1347	715	419	516



ETA Monitooring; Media monitoring at Meedius

Media monitoring



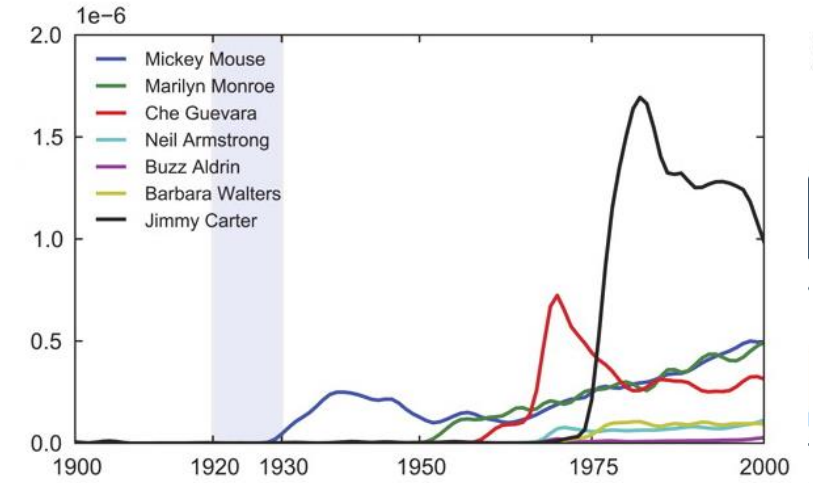
Erakondade kajastatus meedias

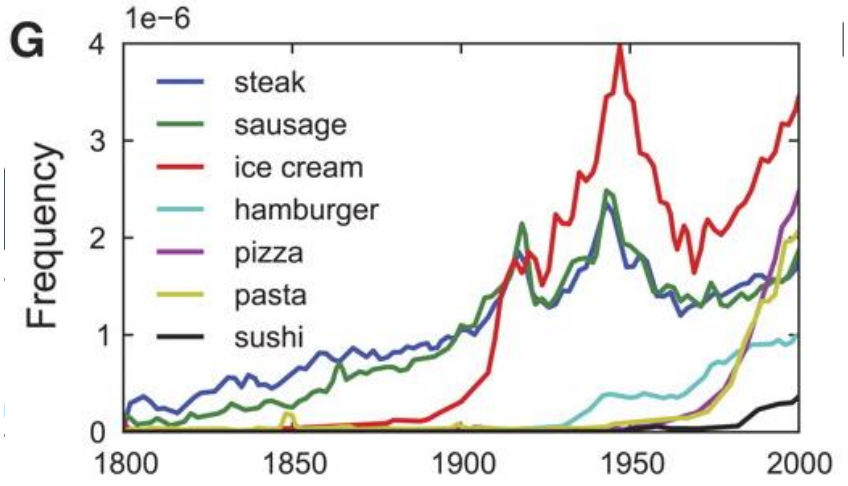
Kui palju on erakondi sel kuul (eelmisel kuul) uudistes mainitud?

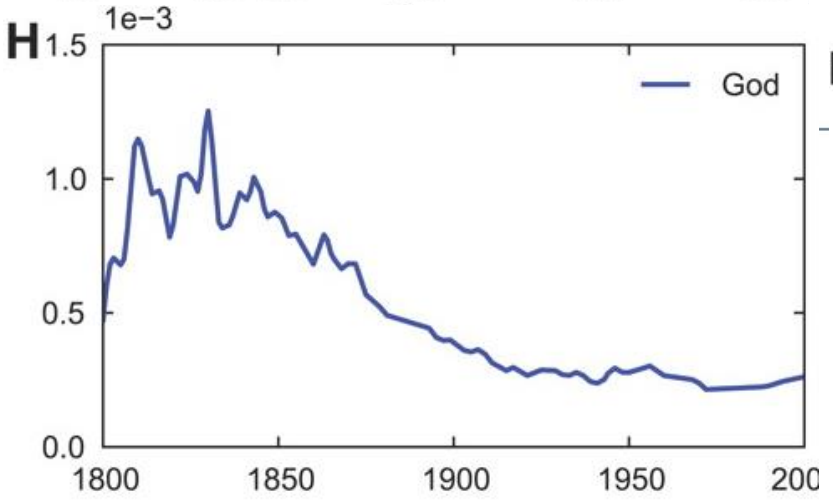
Erakond	Mainimisi kokku	Negatiivseid	Positiivseid
Reformierakond	145 (157)	33 (28)	33 (41)
KESKERAKOND	319 (288)	62 (50)	77 (65)
IRL	101 (116)	24 (26)	22 (34)
SDE	217 (185)	34 (38)	52 (41)

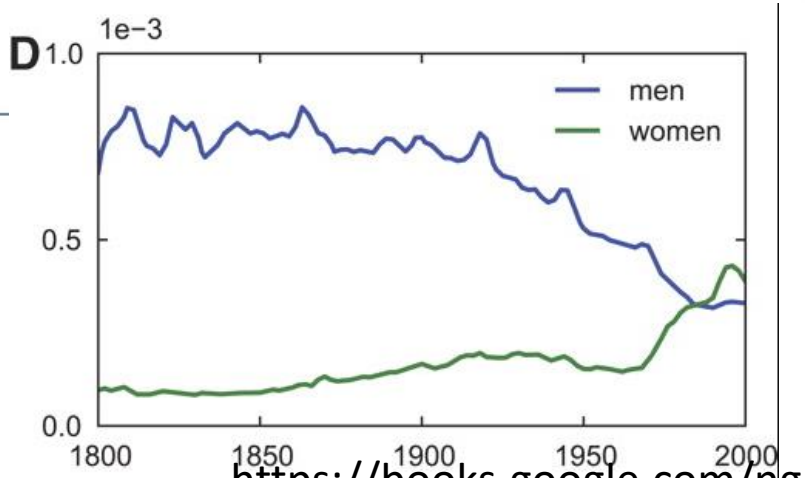
http://meediainspektor.ee/ (student's work with simple tools)

„All digitized texts“





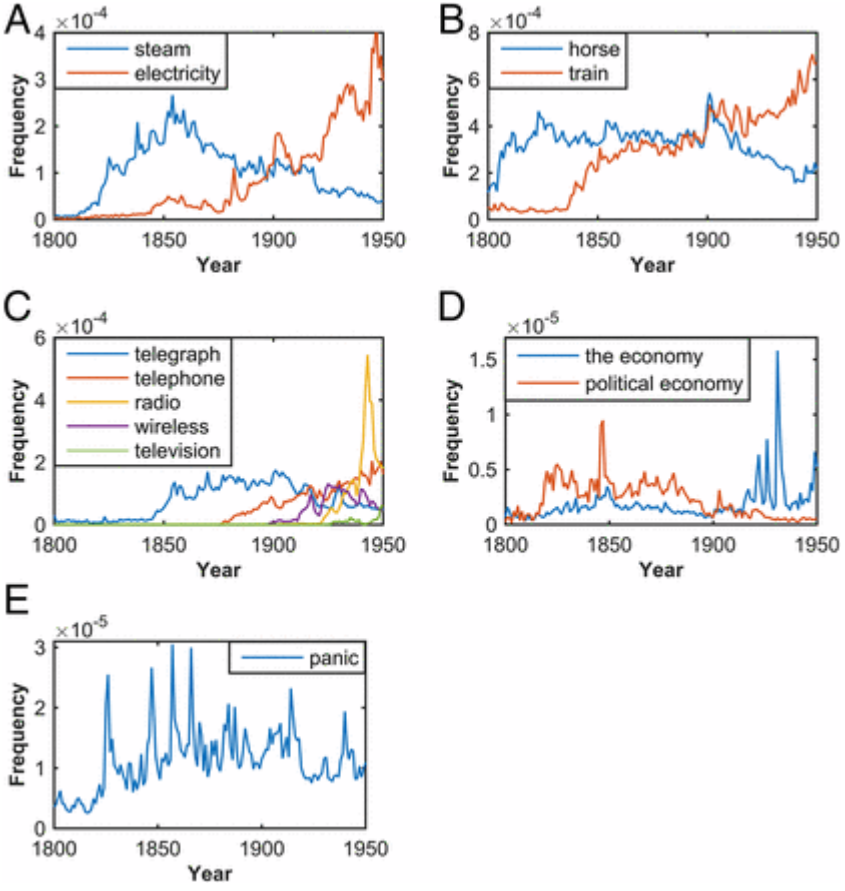


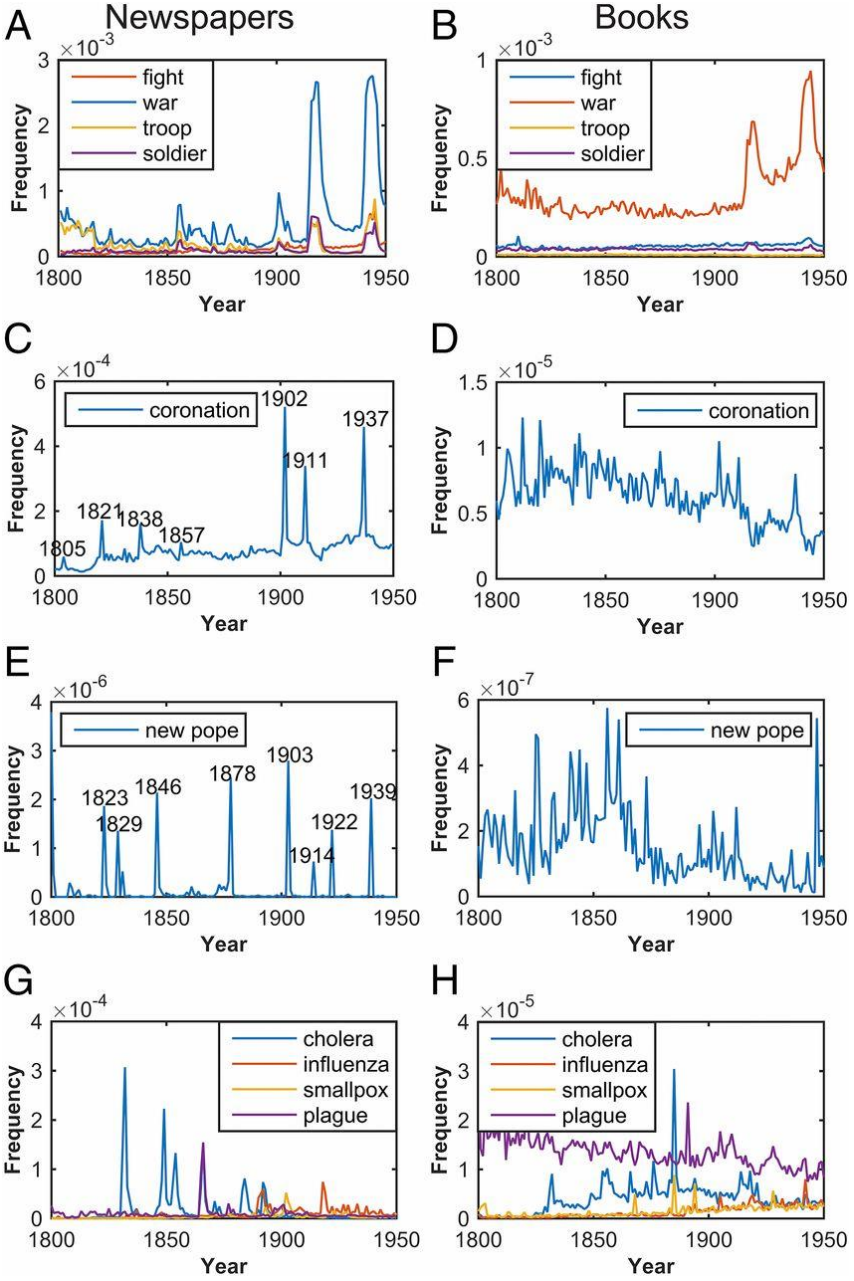


https://books.google.com/ngrams/

Michel et al. 2008 Quantitative Analysis of Culture Using Millions of Digitized Books

150 years of news





Lansdall-Welfare 2017. Content analysis of 150 years of British periodicals

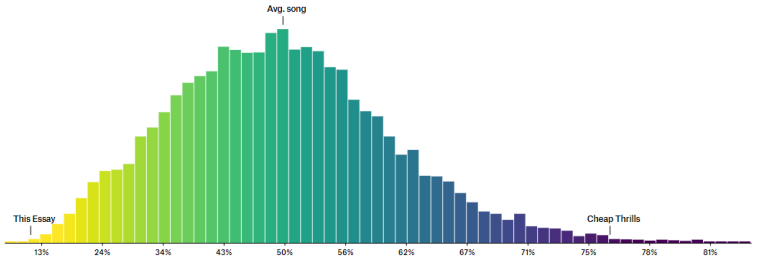
Are Pop Lyrics Getting More Repetitive?

- 15 000 songs of top 100-s 1957-2017
- .zip the text

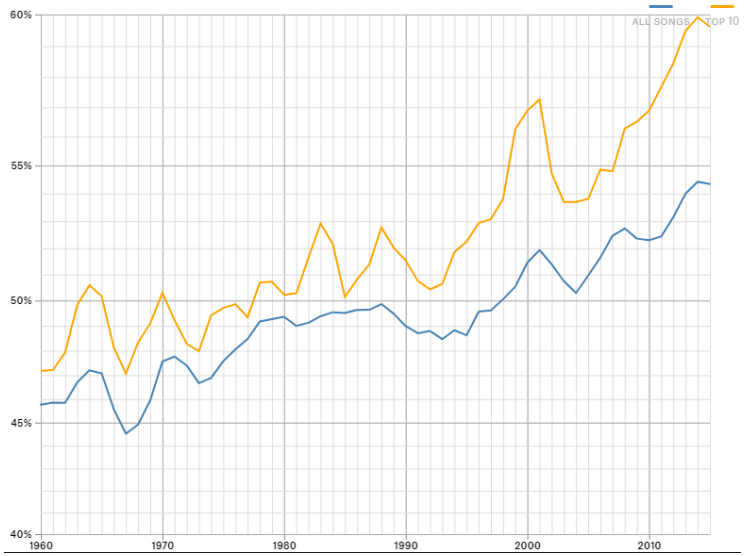
```
baby I don't need dollar bills to have fun tonight
I love cheap thrills!
baby I don't need dollar bills to have fun tonight
I love cheap thrills!
I don't need no money
as long as I can feel the beat
I don't need no money
as long as I keep dancing
```

```
tonight I need dollar bills
I don't keep fun
cheap thr●●s long to feel money
the b●●●s d●n't n●●d t●e dancing baby
fun d●●lar dan●ing thr●lls t●e b●by I need
d●●t have f●n
no no don't●have dan●ing f●n ton●ght
beat ●e can as I d●n't f●el thrills
love the dancing money
```

29.5% Size Reduction

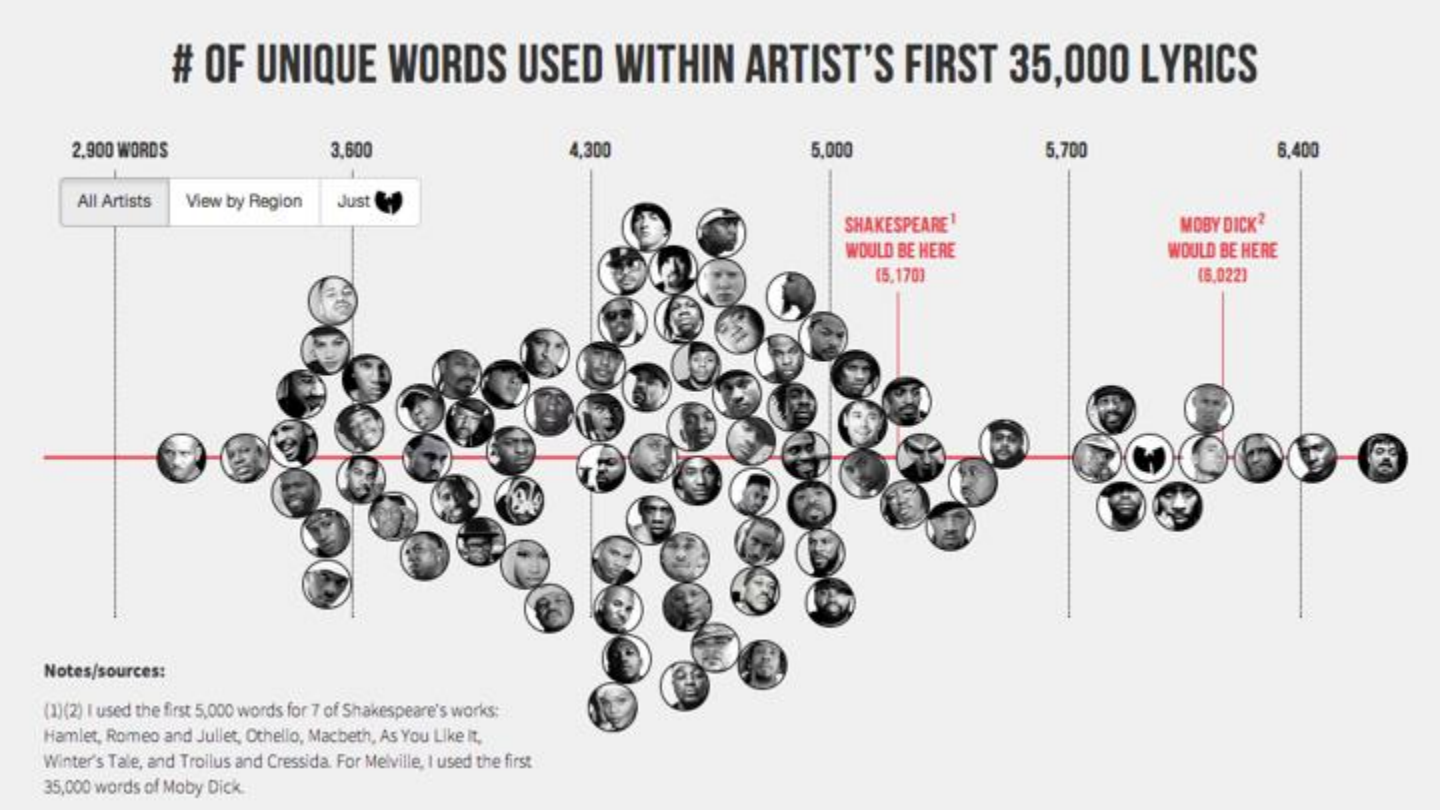


Repetition of Popular Music, by Year



https://pudding.cool/2017/05/song-repetition/

Vocabulary of rap artists

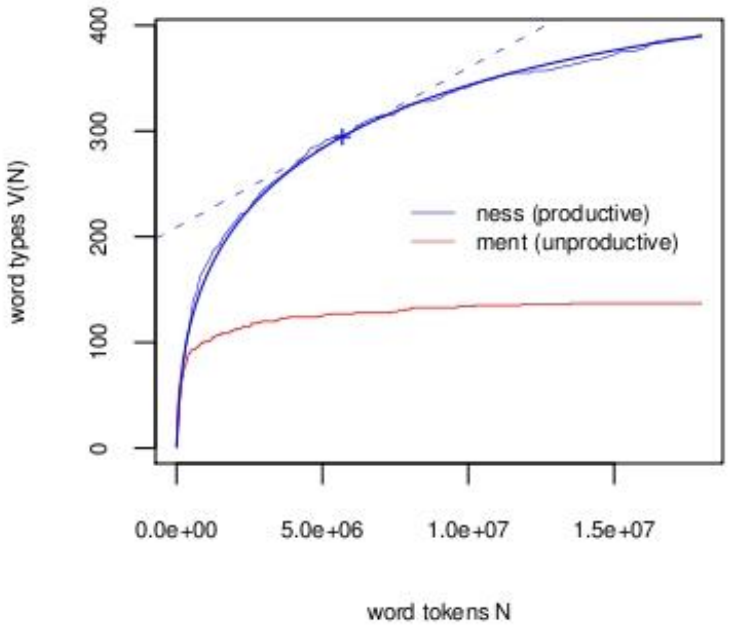


https://pudding.cool/projects/vocabulary/

Morphological productivity

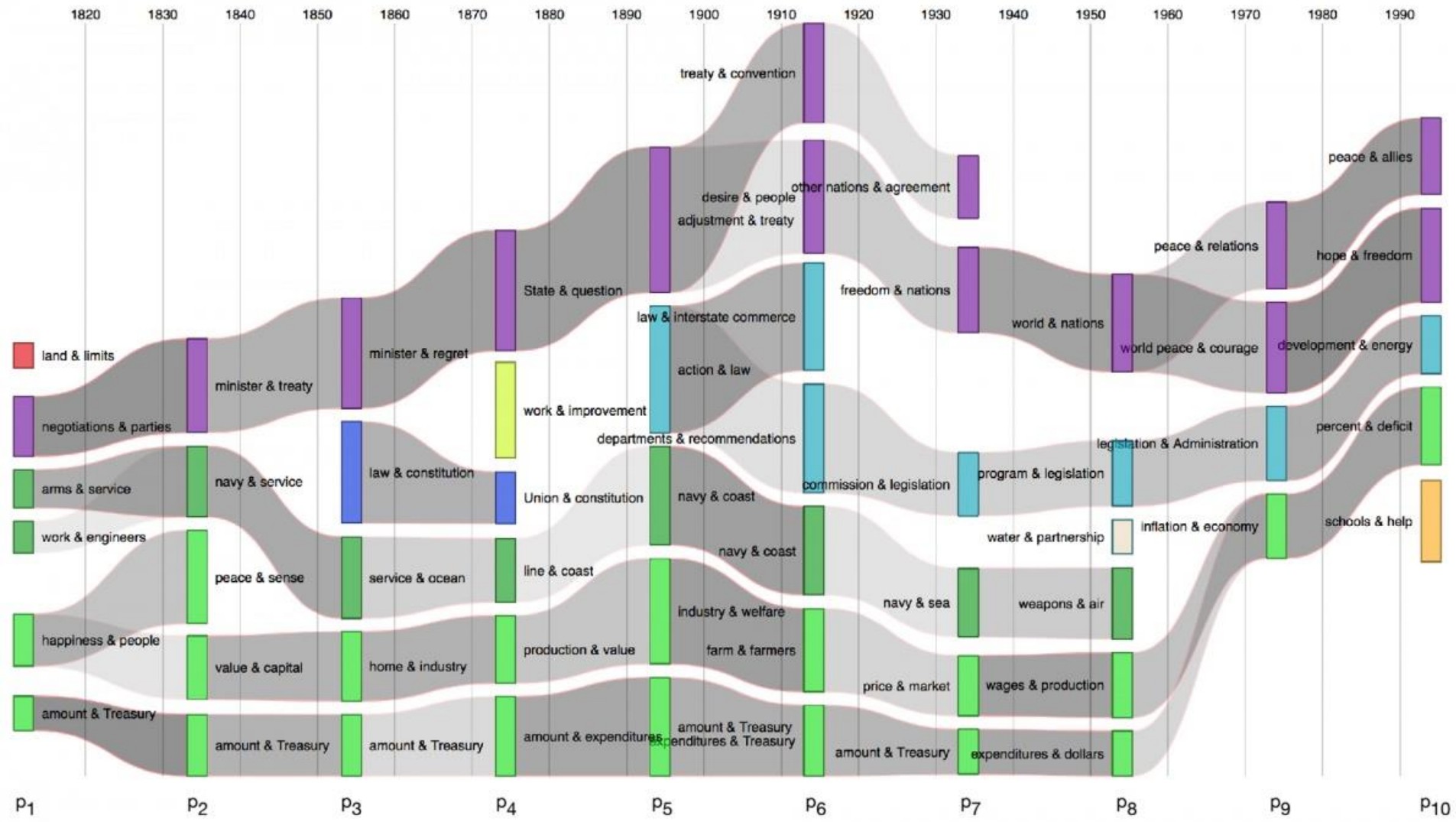
- Number of types / number of tokens

- '-ness' – (*goodness)*
- '-ment' – (*improvement)*



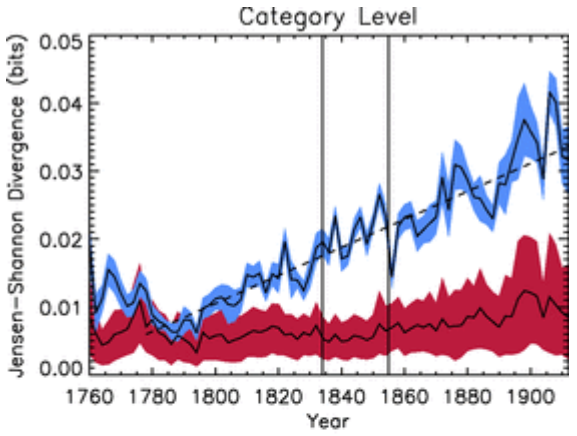
Baayen 1992. Quantitative aspects of morphological productivity

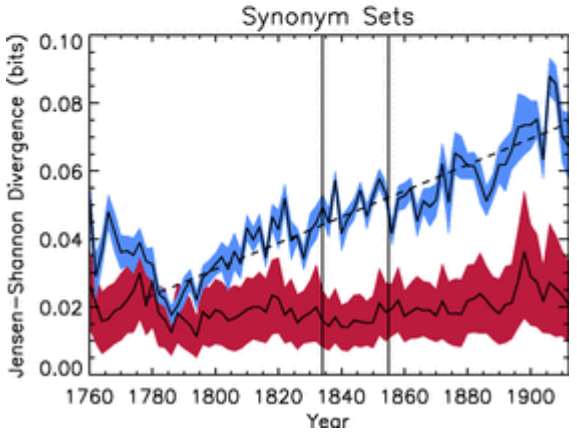
Presidential speechs in USA



Rule et al. 2015 Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014

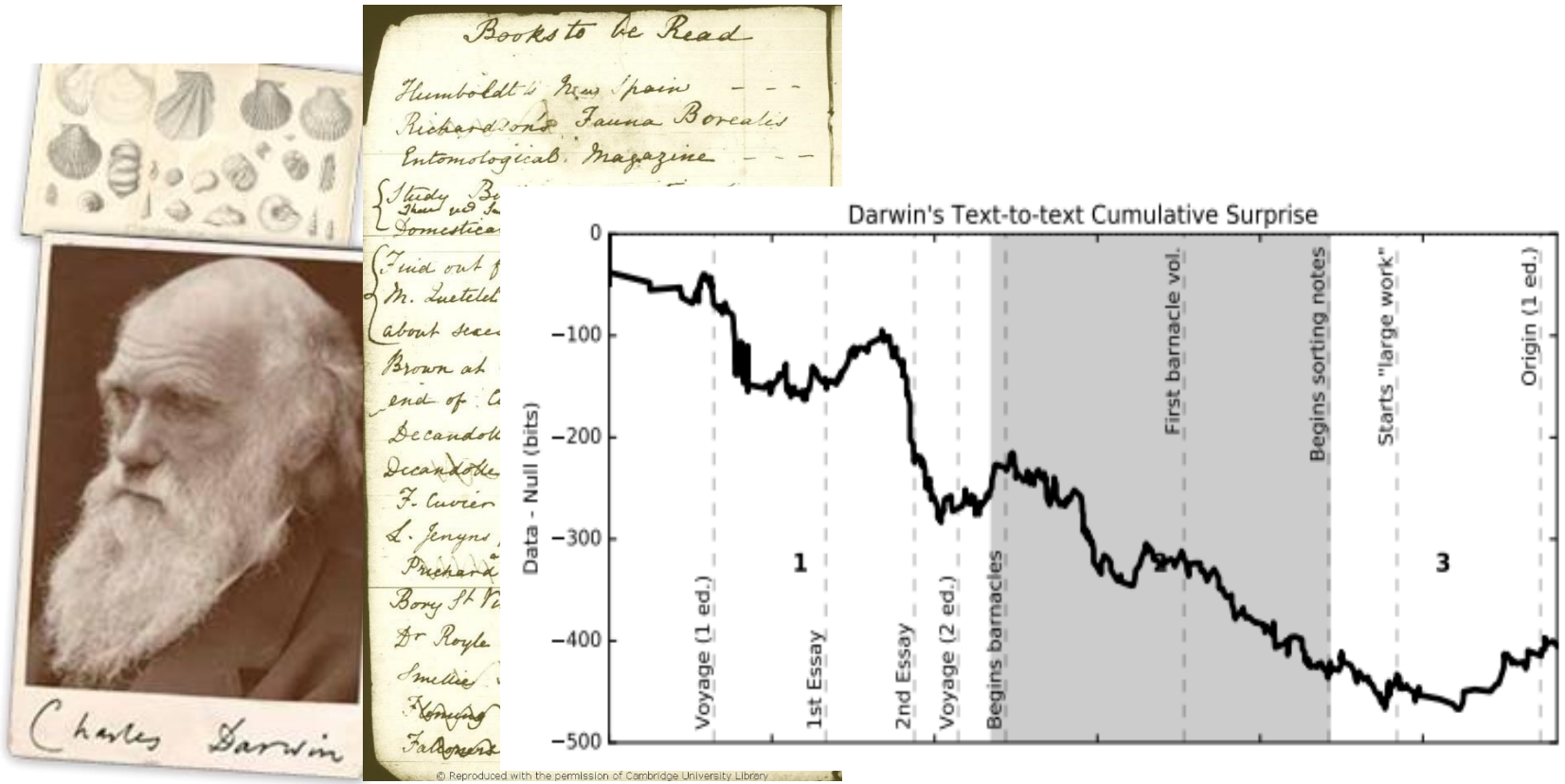
Civilizing process





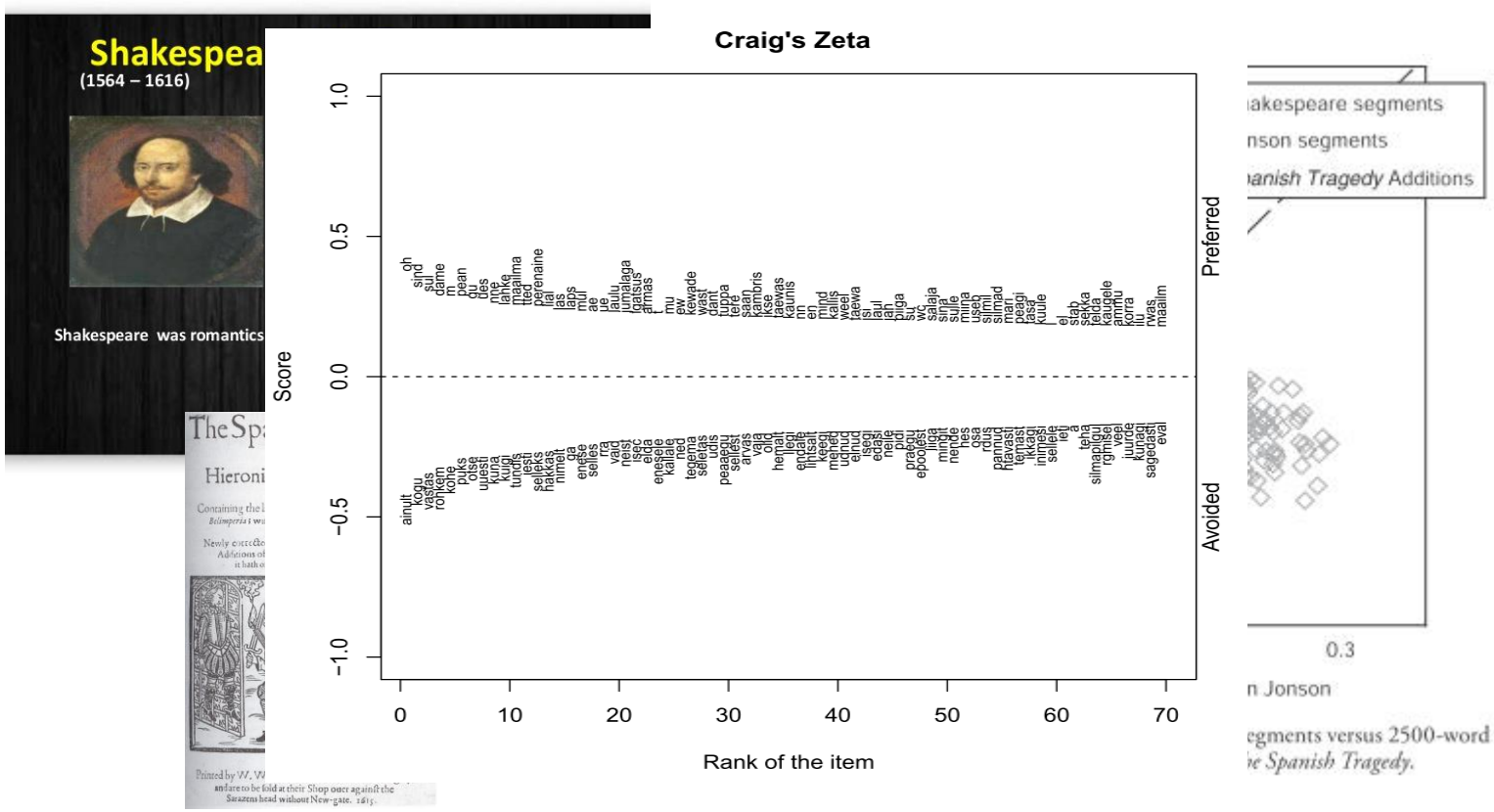
Klingenstein et al. 2014. The civilizing process in London's Old Bailey

Darwin's reading habits



Murdock et al. 2017. Exploration and exploitation of Victorian science in Darwin's reading notebooks.

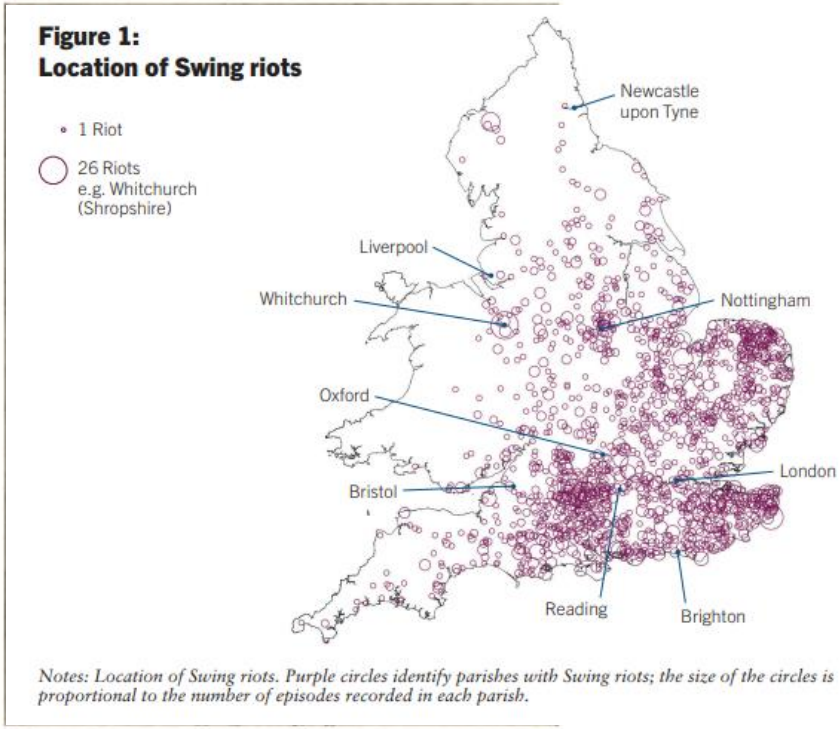
Finding the author

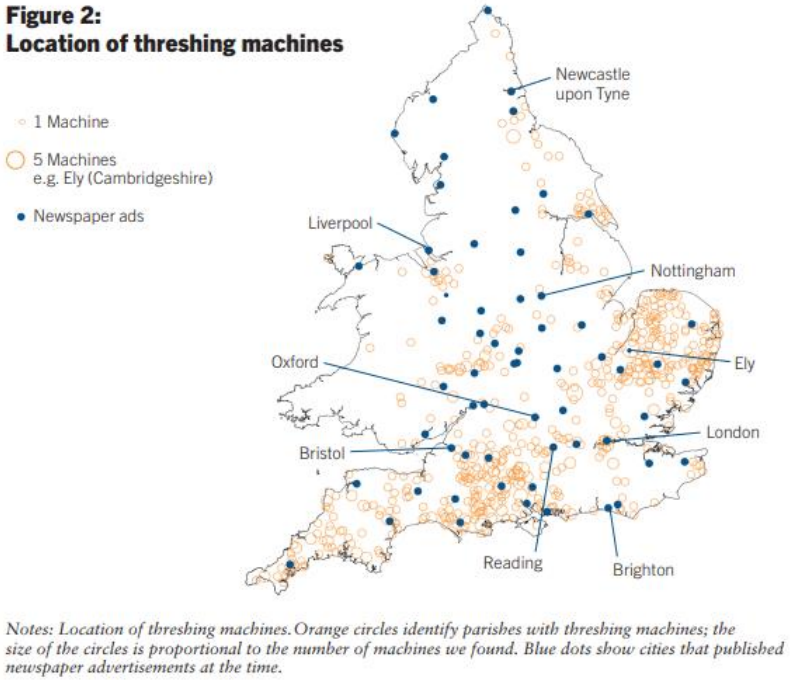


Craig & Kinney 2009. Shakespeare, Computers, and the Mystery of Authorship

Deep mining history

- Advertisements in newspapers -> diffusion of technology
- „Swing riots“ 1829-1830: machines, unemployment and crops.

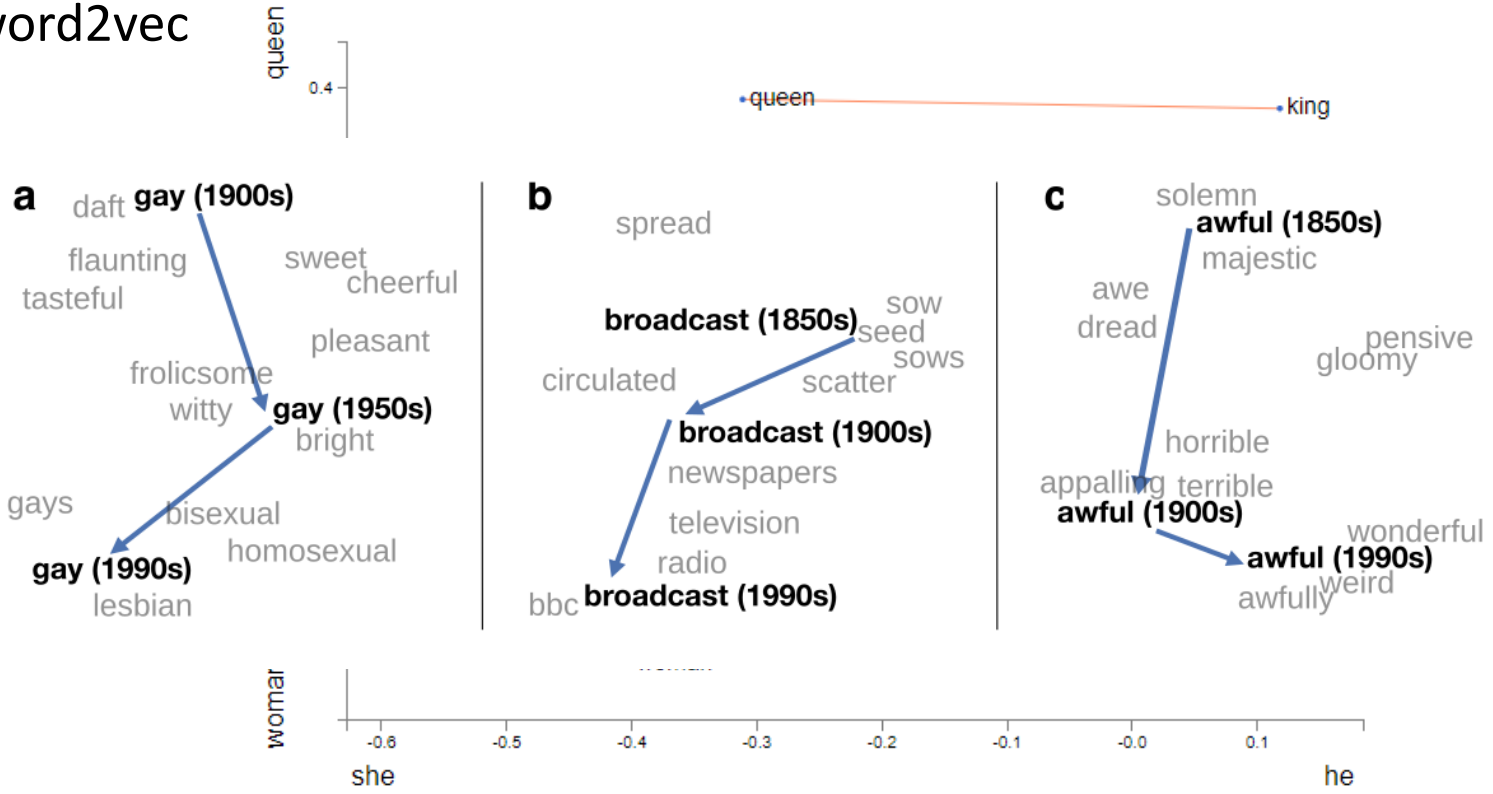




Caprettini & Voth 2018 Rage against the machines: New technology and violent unrest in industrializing England

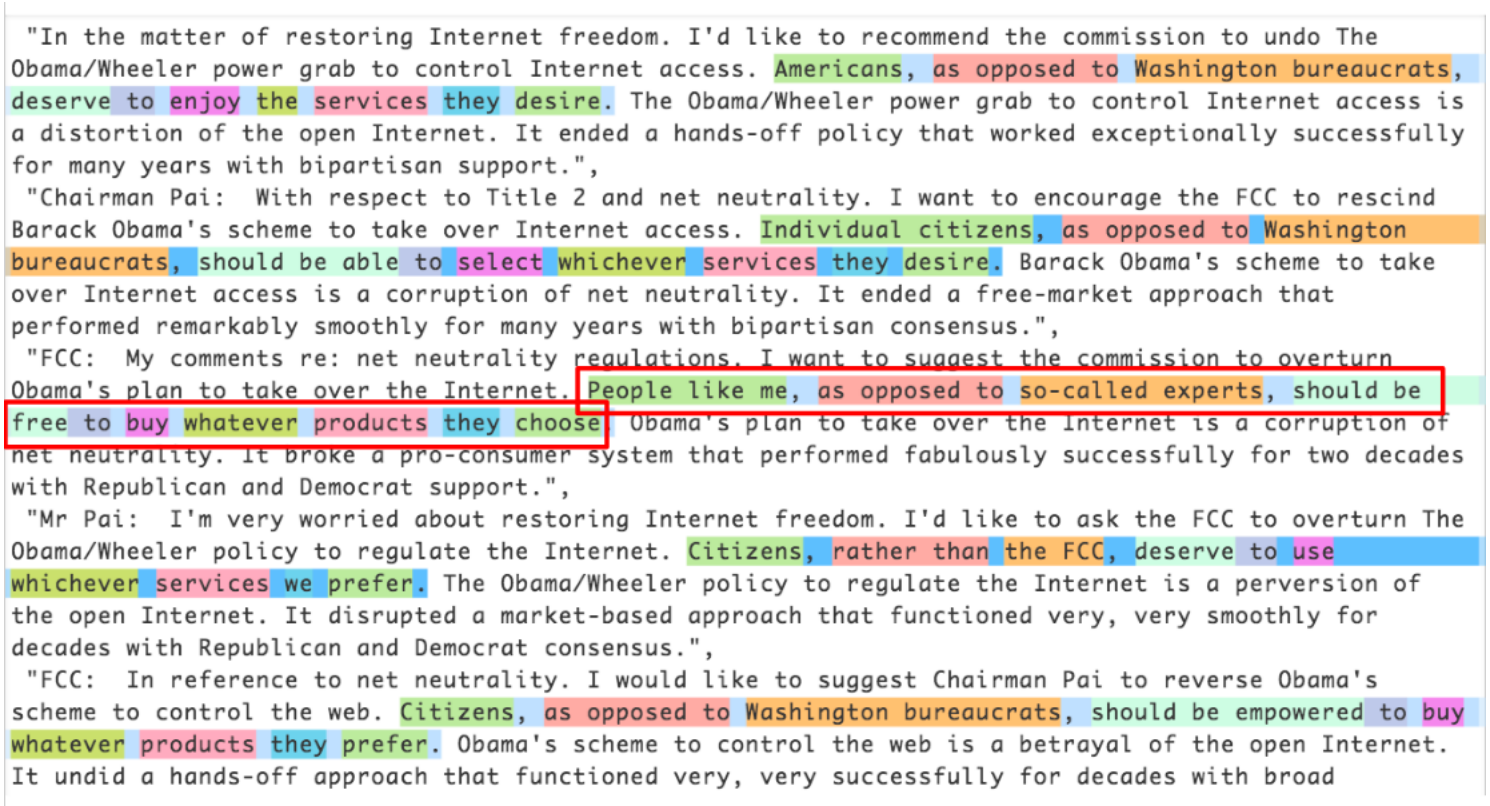
Word semantics

- word2vec



https://lamyiowce.github.io/word2viz/

Plagiarism and bot detection



"In the matter of restoring Internet freedom. I'd like to recommend the commission to undo The Obama/Wheeler power grab to control Internet access. Americans, as opposed to Washington bureaucrats, deserve to enjoy the services they desire. The Obama/Wheeler power grab to control Internet access is a distortion of the open Internet. It ended a hands-off policy that worked exceptionally successfully for many years with bipartisan support.",

"Chairman Pai: With respect to Title 2 and net neutrality. I want to encourage the FCC to rescind Barack Obama's scheme to take over Internet access. Individual citizens, as opposed to Washington bureaucrats, should be able to select whichever services they desire. Barack Obama's scheme to take over Internet access is a corruption of net neutrality. It ended a free-market approach that performed remarkably smoothly for many years with bipartisan consensus.",

"FCC: My comments re: net neutrality regulations. I want to suggest the commission to overturn Obama's plan to take over the Internet. People like me, as opposed to so-called experts, should be free to buy whatever products they choose. Obama's plan to take over the Internet is a corruption of net neutrality. It broke a pro-consumer system that performed fabulously successfully for two decades with Republican and Democrat support.",

"Mr Pai: I'm very worried about restoring Internet freedom. I'd like to ask the FCC to overturn The Obama/Wheeler policy to regulate the Internet. Citizens, rather than the FCC, deserve to use whichever services we prefer. The Obama/Wheeler policy to regulate the Internet is a perversion of the open Internet. It disrupted a market-based approach that functioned very, very smoothly for decades with Republican and Democrat consensus.",

"FCC: In reference to net neutrality. I would like to suggest Chairman Pai to reverse Obama's scheme to control the web. Citizens, as opposed to Washington bureaucrats, should be empowered to buy whatever products they prefer. Obama's scheme to control the web is a betrayal of the open Internet. It undid a hands-off approach that functioned very, very successfully for decades with broad

Kao 2017. More than a Million Pro-Repeal Net Neutrality Comments were Likely Faked

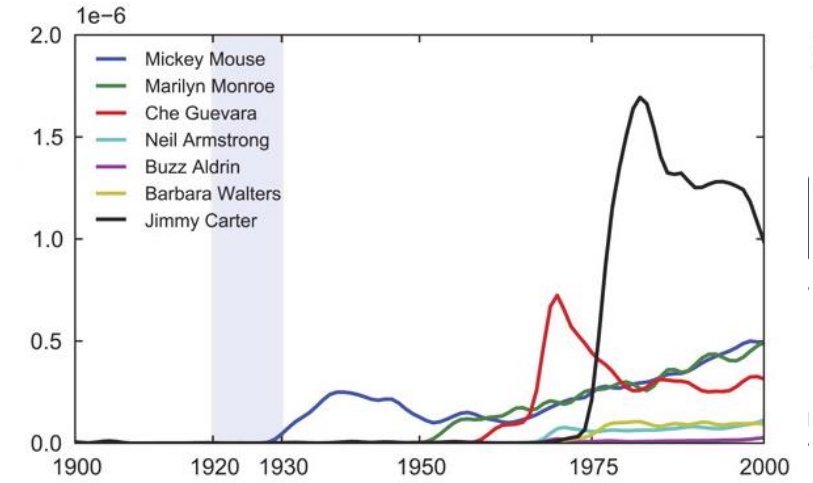
Basic methods

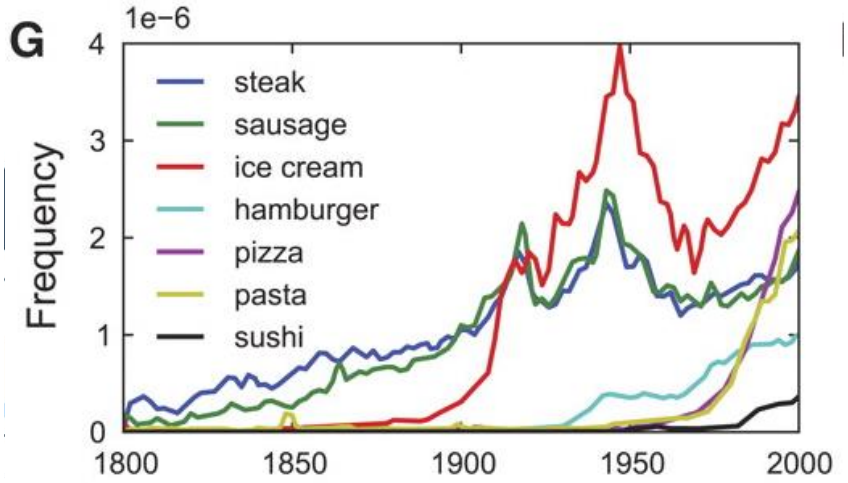
Basic methods

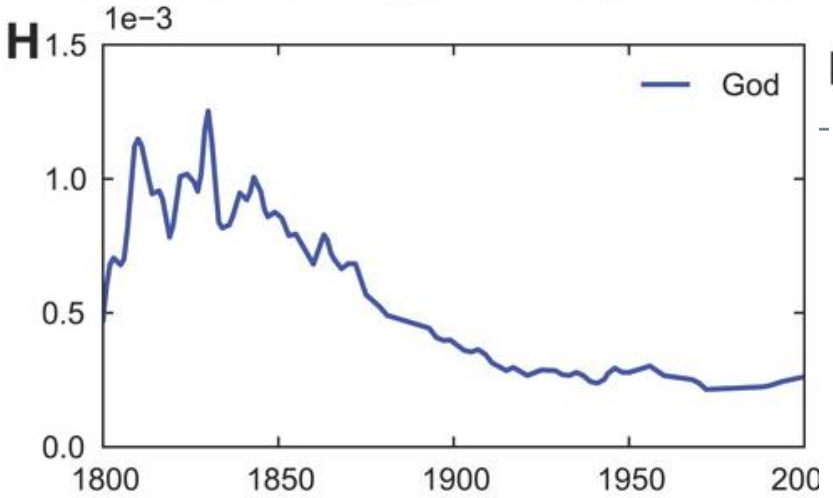
- 1. Frequencies
- 2. Comparing texts
- 3. Text-internal structure
- 4. Lexicon-based
- 5. Word distributions
- 6. Metainformation!

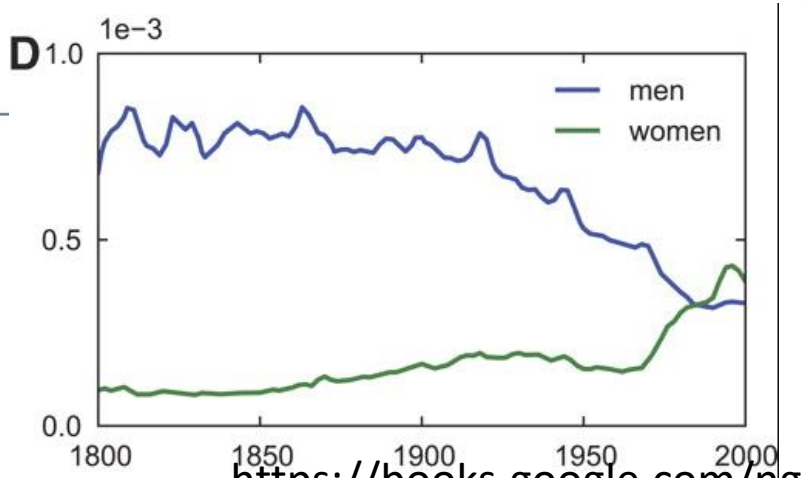
Frequencies

Frequencies





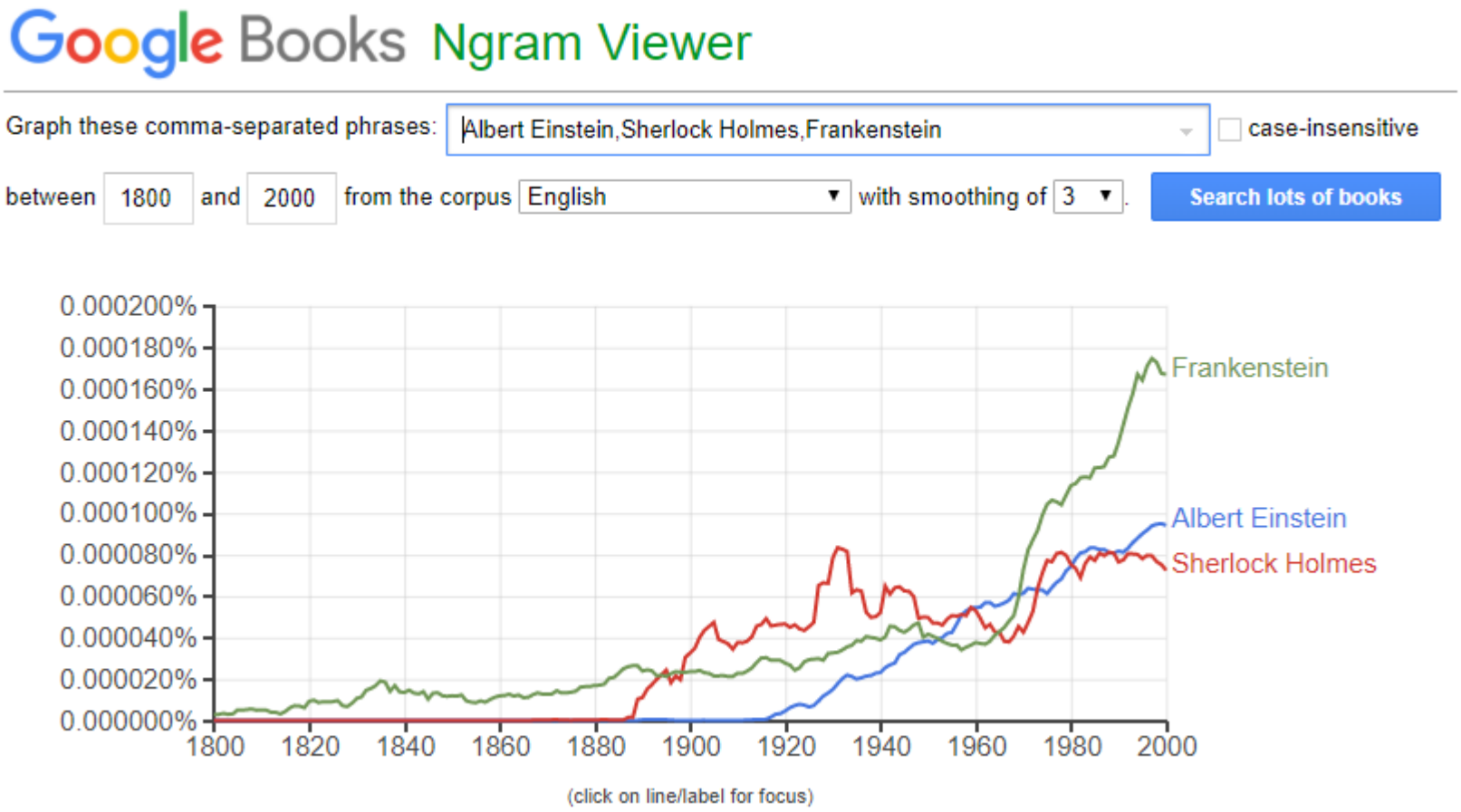




https://books.google.com/ngrams/

Michel et al. 2008 Quantitative Analysis of Culture Using Millions of Digitized Books

Google Ngrams
(1-grams, 2-grams, 3-grams...)



https://books.google.com/ngrams

Counts vs frequencies

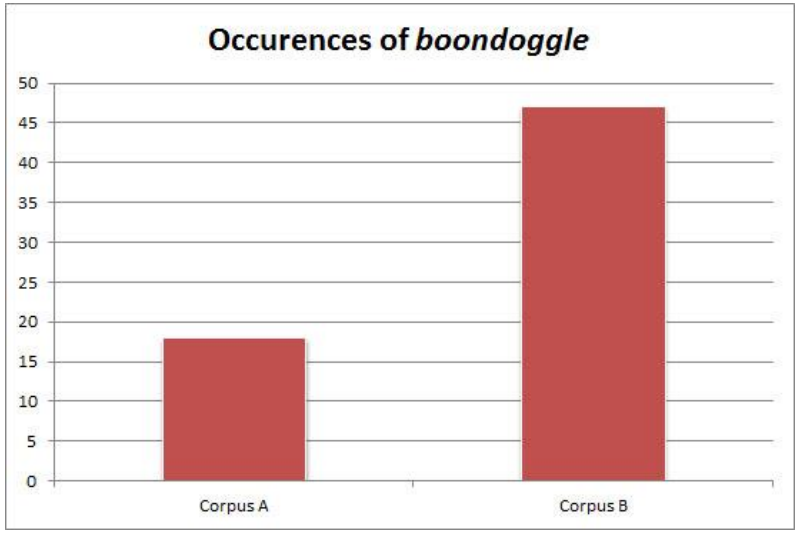
- The 1-gram ‘boondoggle’ in corpus
- (A **boondoggle** is a project that is considered a waste of both time and money, yet is often continued due to extraneous policy or political motivations.)

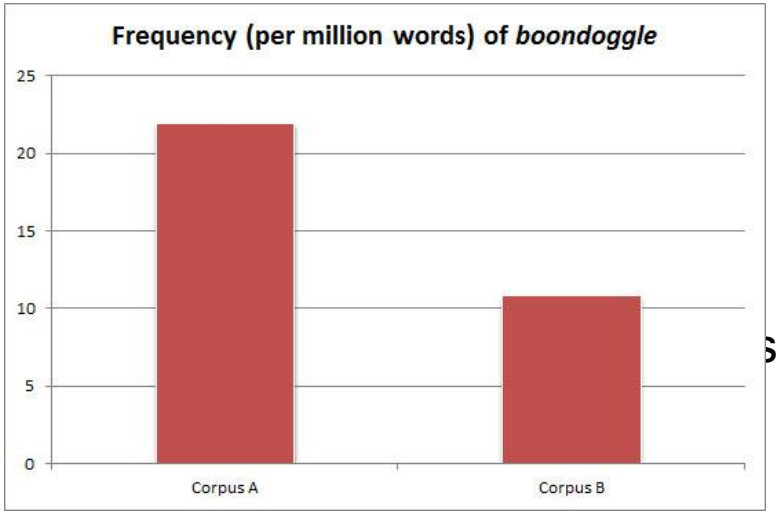
$$\frac{18}{821{,}273} = \frac{x}{1{,}000{,}000}$$

$$x\,(821{,}273) = 18\,(1{,}000{,}000)$$

$$x = 18\,(1{,}000{,}000) / (821{,}273)$$

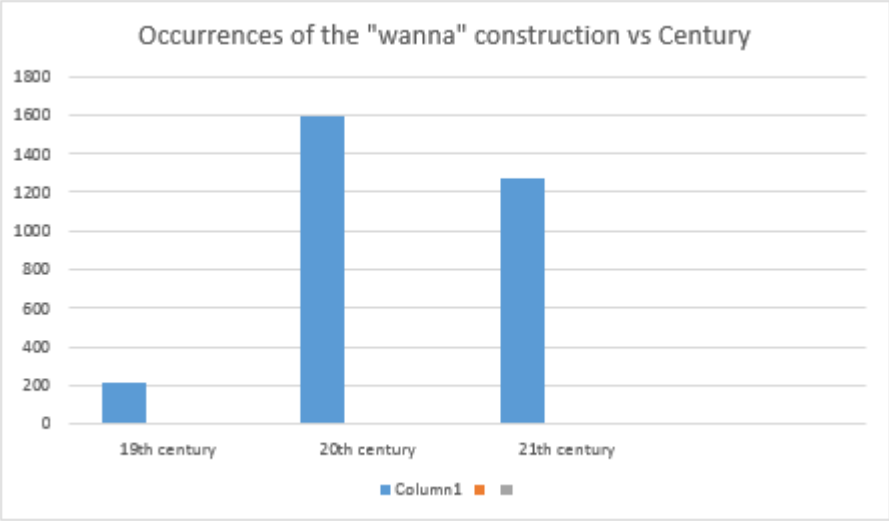
$$x = \sim 22$$





http://www.thegrammarlab.com/?p=160

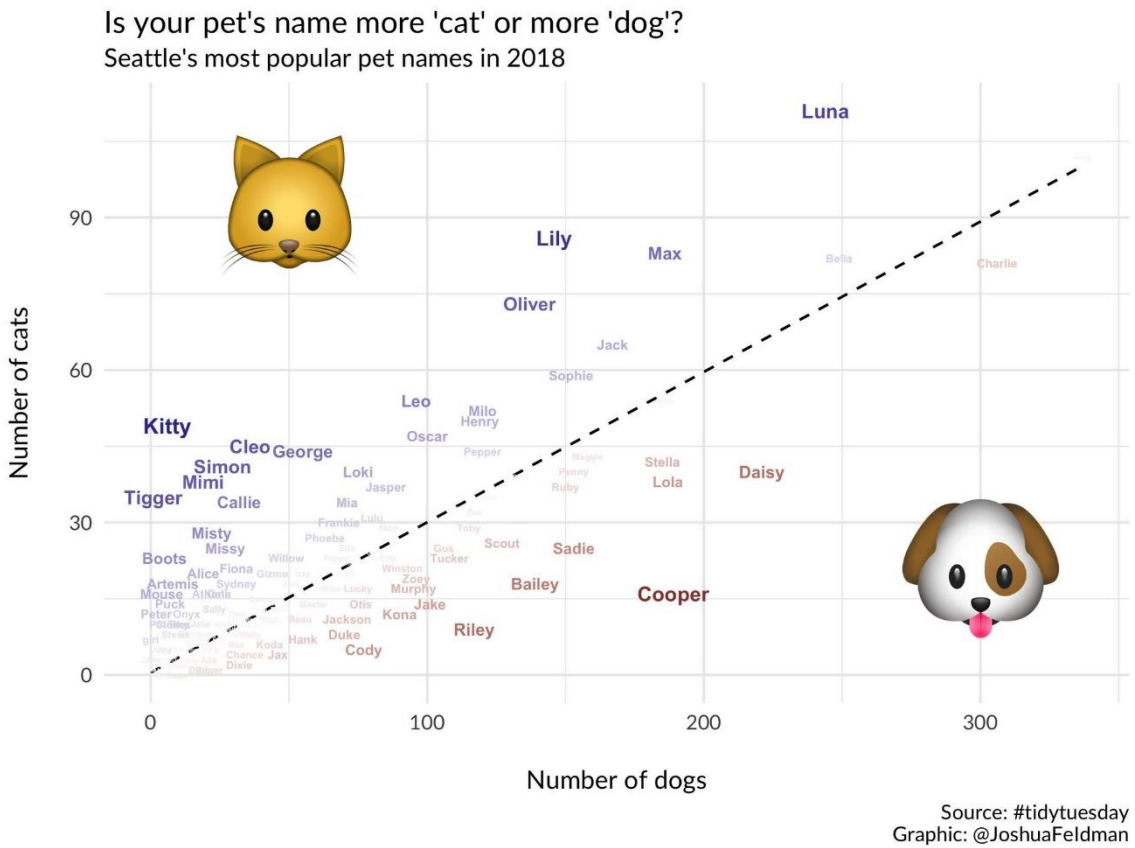
Quick test of attention



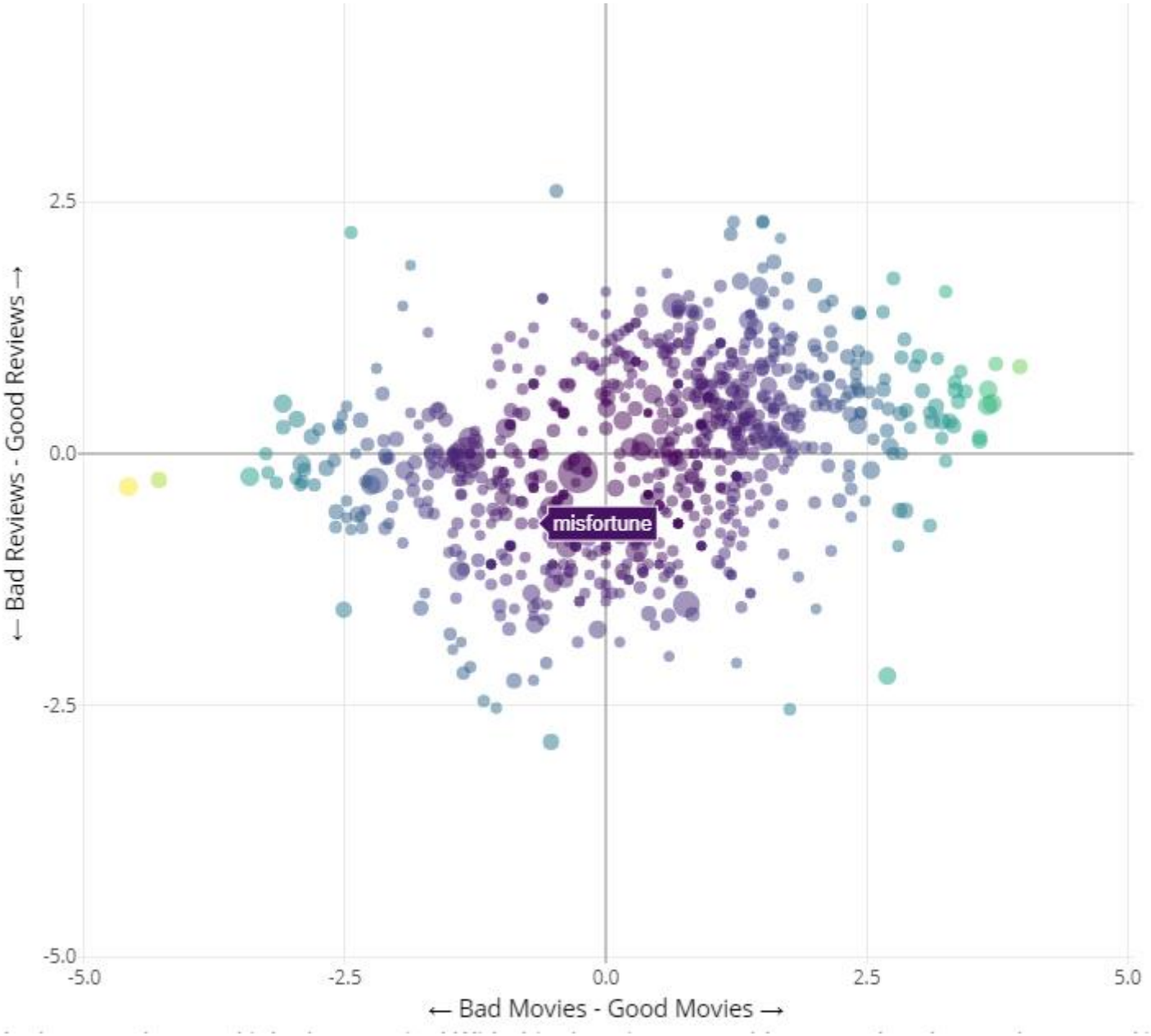
[StackExchange: How do I compare data from two corpora correctly?]

Comparing texts

Comparing frequencies

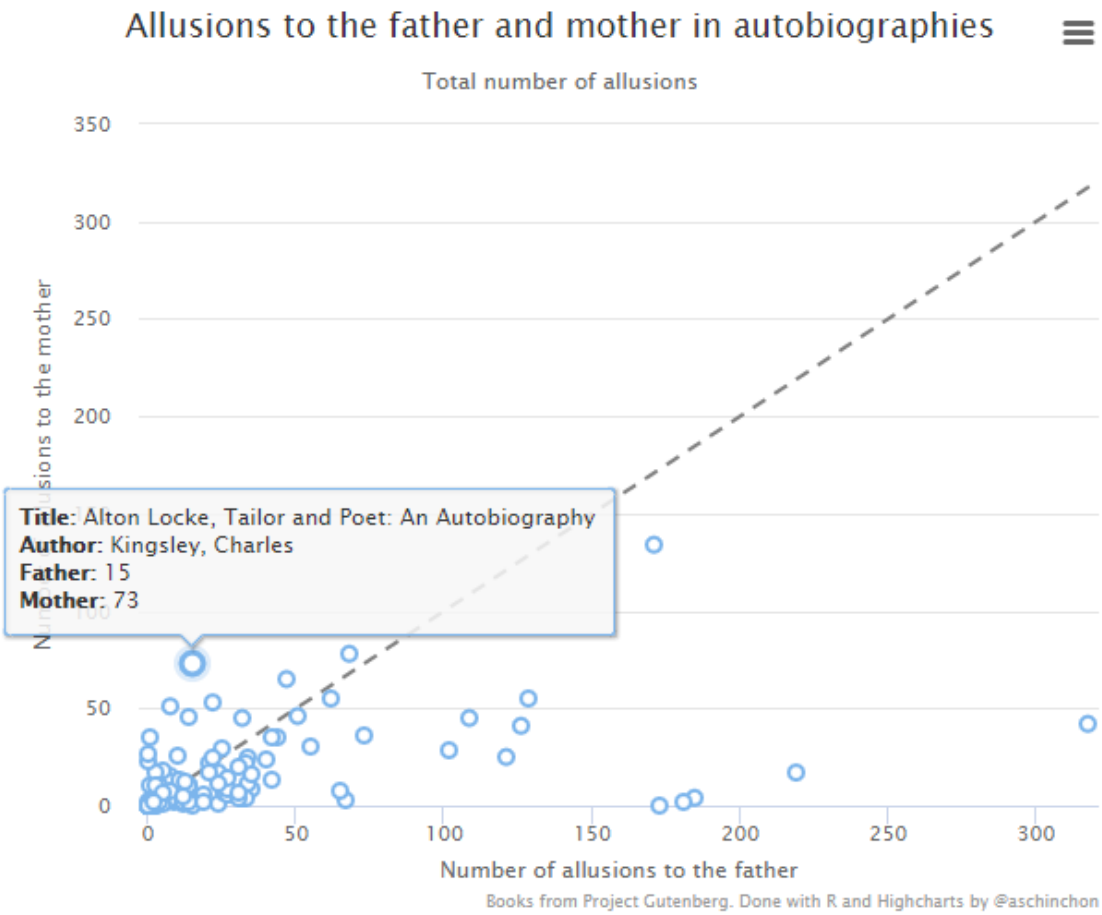


https://twitter.com/JoshuaFeldman/status/1110604459669897218



https://www.hvitfeldt.me/blog/what-are-the-reviews-telling-us/

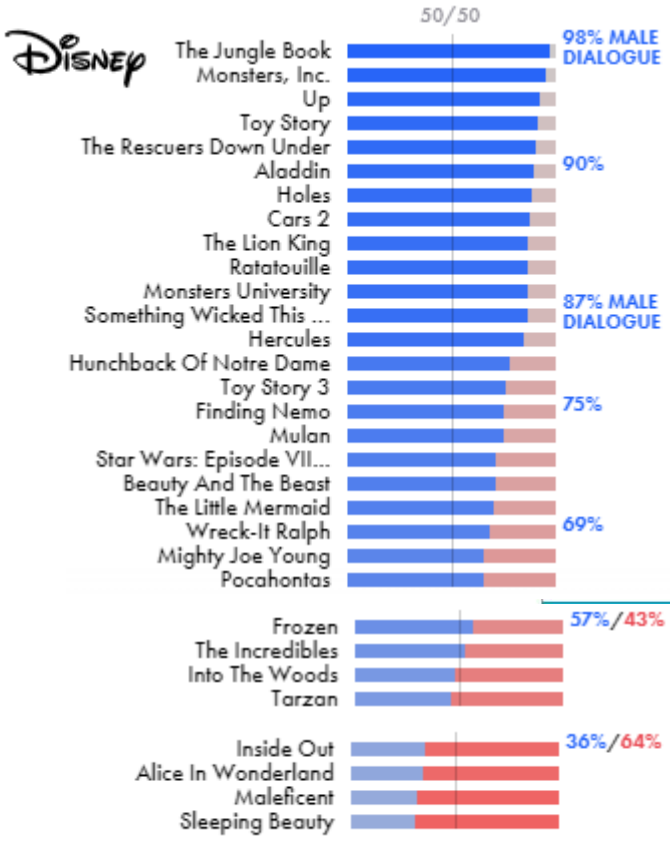
Lexicon+comparison

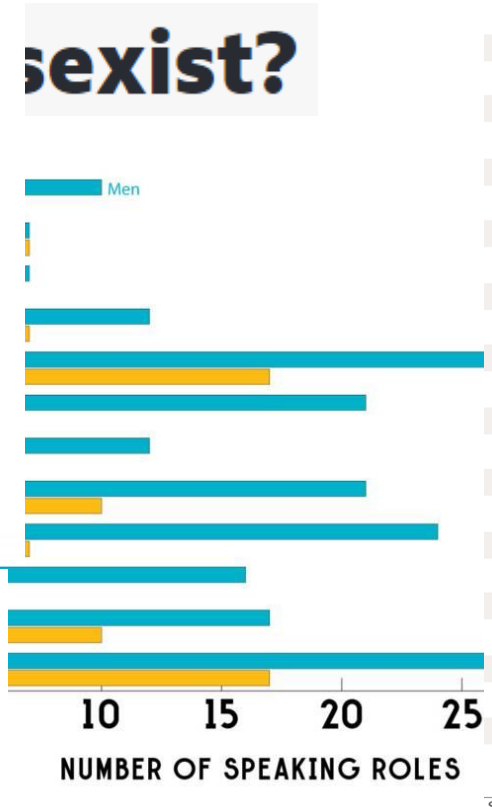


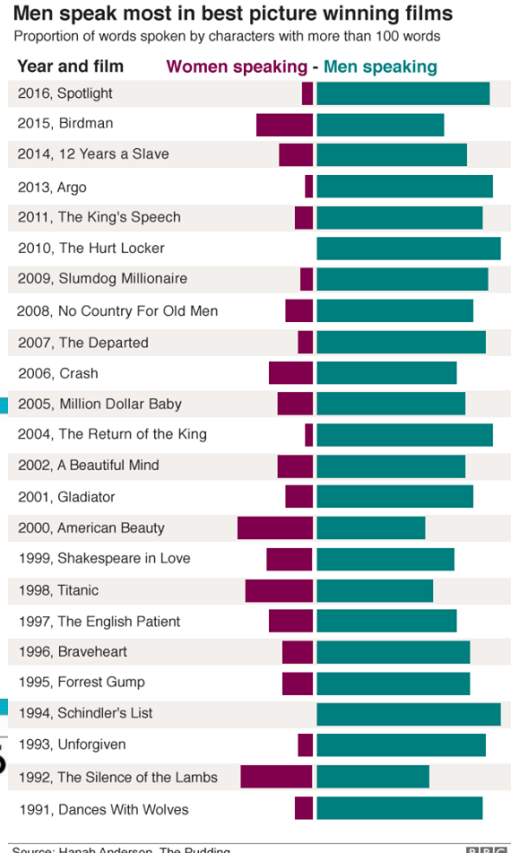
https://fronkonstin.com/2019/02/24/allusions-to-parents-in-autobiographies-or-reading-118-books-in-a-few-seconds/

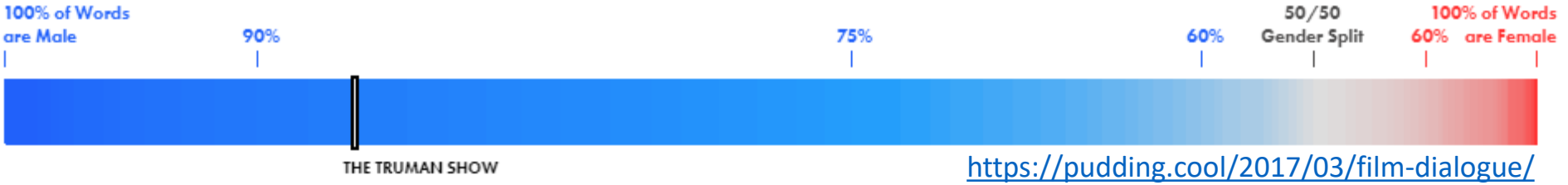
Text-internal structure

Who speaks when



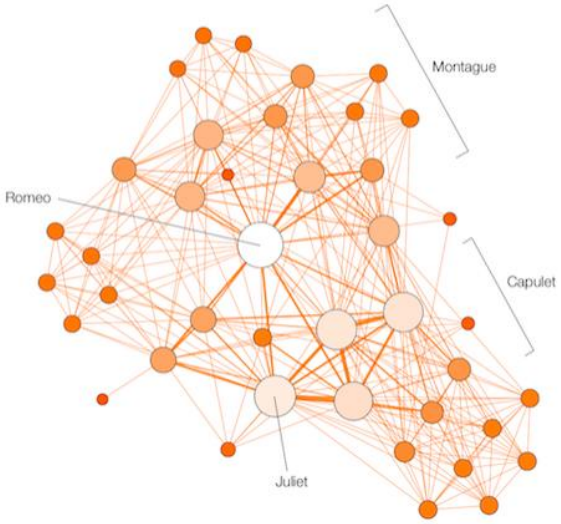






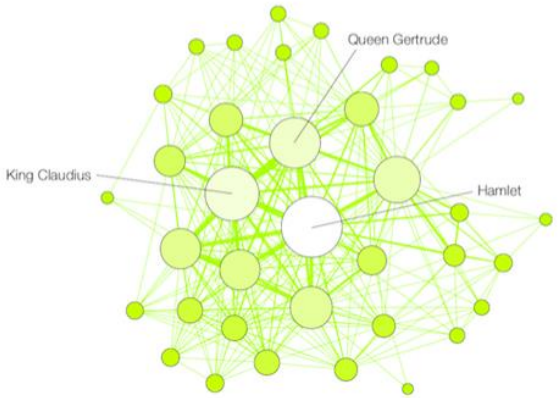
https://pudding.cool/2017/03/film-dialogue/

Co-occurrences in scenes



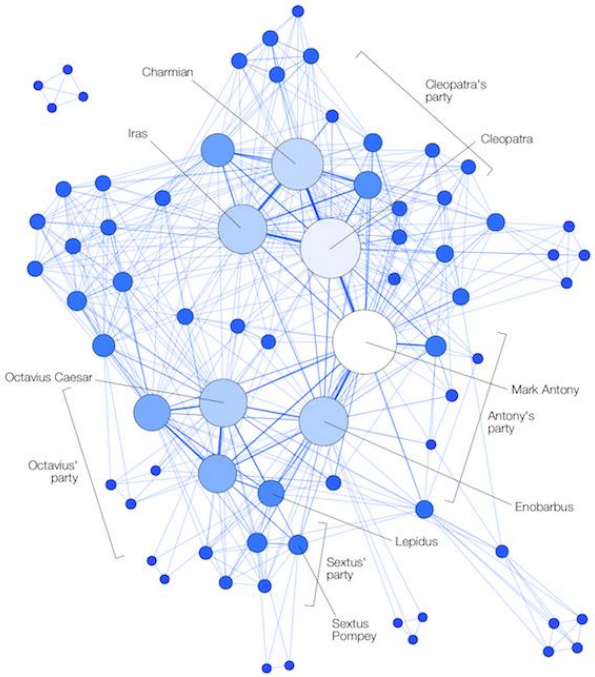
ROMEO AND JULIET

Number of characters 41 | 37% Network density



HAMLET

Number of characters 37 | 39% Network density



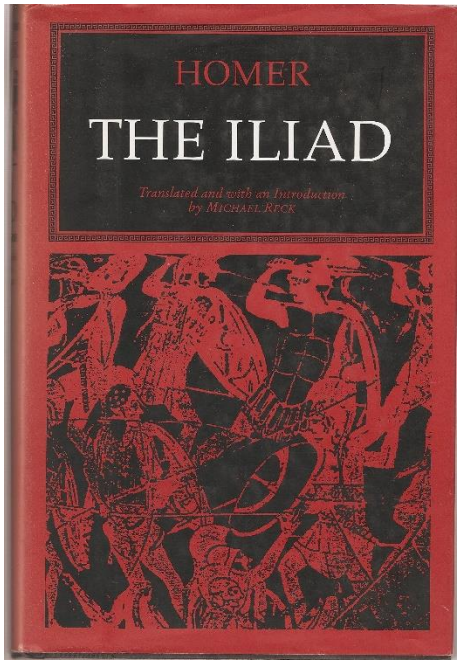
ANTONY AND CLEOPATRA

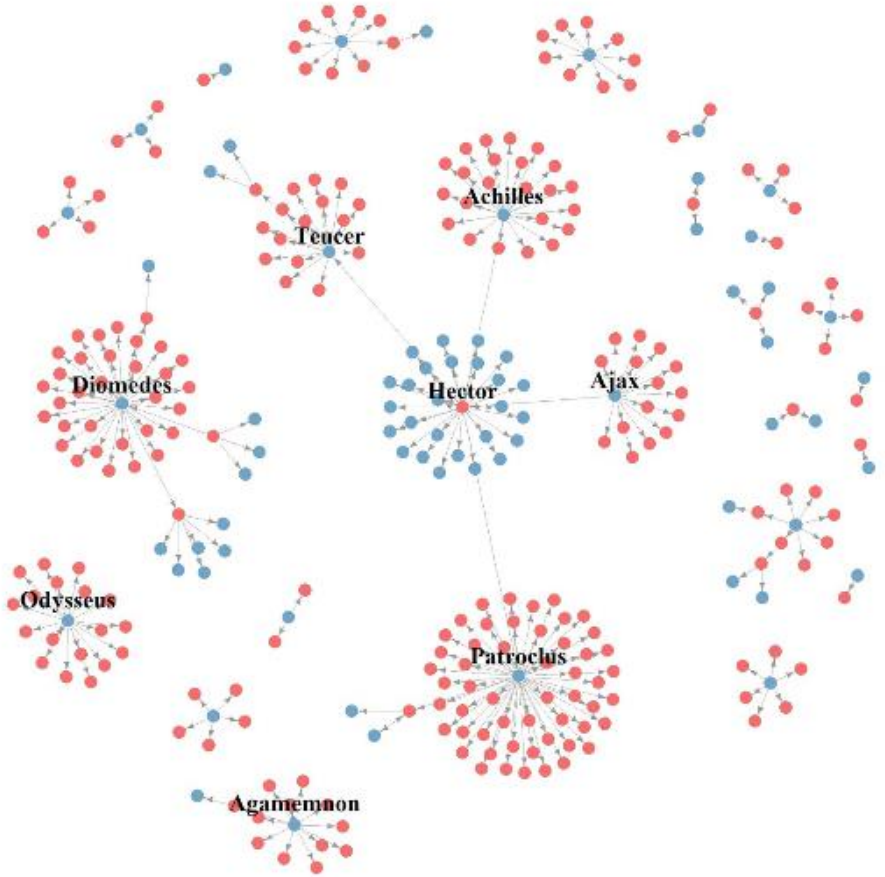
Number of characters 74 | 17% Network density

Grandjean 2015. Network visualization: mapping Shakespeare's tragedies

Cooccurrences in scenes

- Fights in Iliad



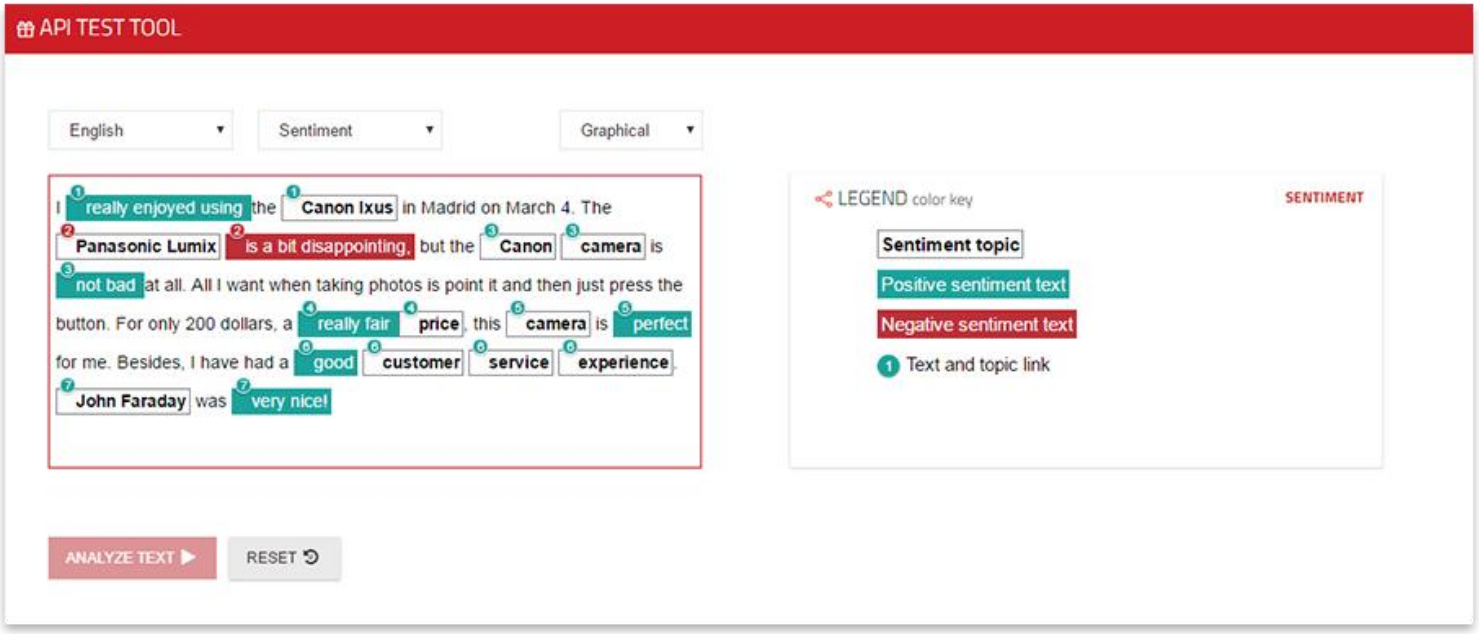


Rossmann 2015. Spears and Prisms: Combat Victories in the Iliad

Lexicon-based

Lexicon based

- Sentiment analysis

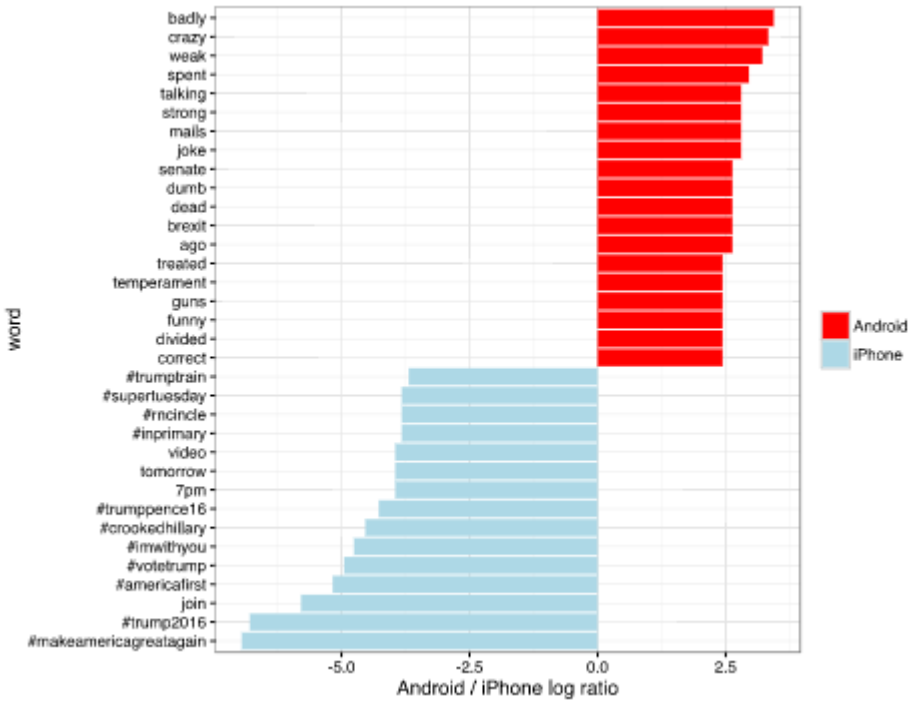


Bitext tool

Political tweets

Text analysis of Trump's tweets confirms he writes only the (angrier) Android half

I don't normally post about politics (I'm not particularly savvy about polling, which is where data science has had the largest impact on politics). But this weekend I saw a hypothesis about Donald Trump's twitter account that simply begged to be investigated with data:

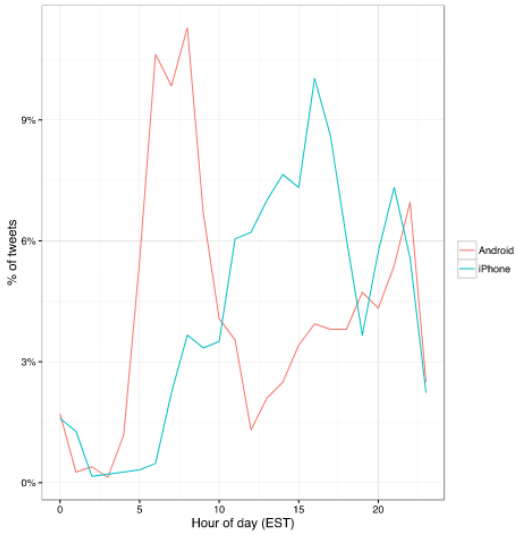


Donald J. Trump
@realDonaldTrump
Follow

Thank you Windham, New Hampshire! #TrumpPence16 #MAGA
5:19 AM - Aug 7, 2016 · Windham, NH
2,338 6,044 19,778

Donald J. Trump
@realDonaldTrump
Follow

The media is going crazy. They totally distort so many things on purpose. Crimea, nuclear, "the baby" and so much more. Very



Robinson 2016. Text analysis of Trump's tweets confirms he writes the (angrier) Android half

Sentiment across texts

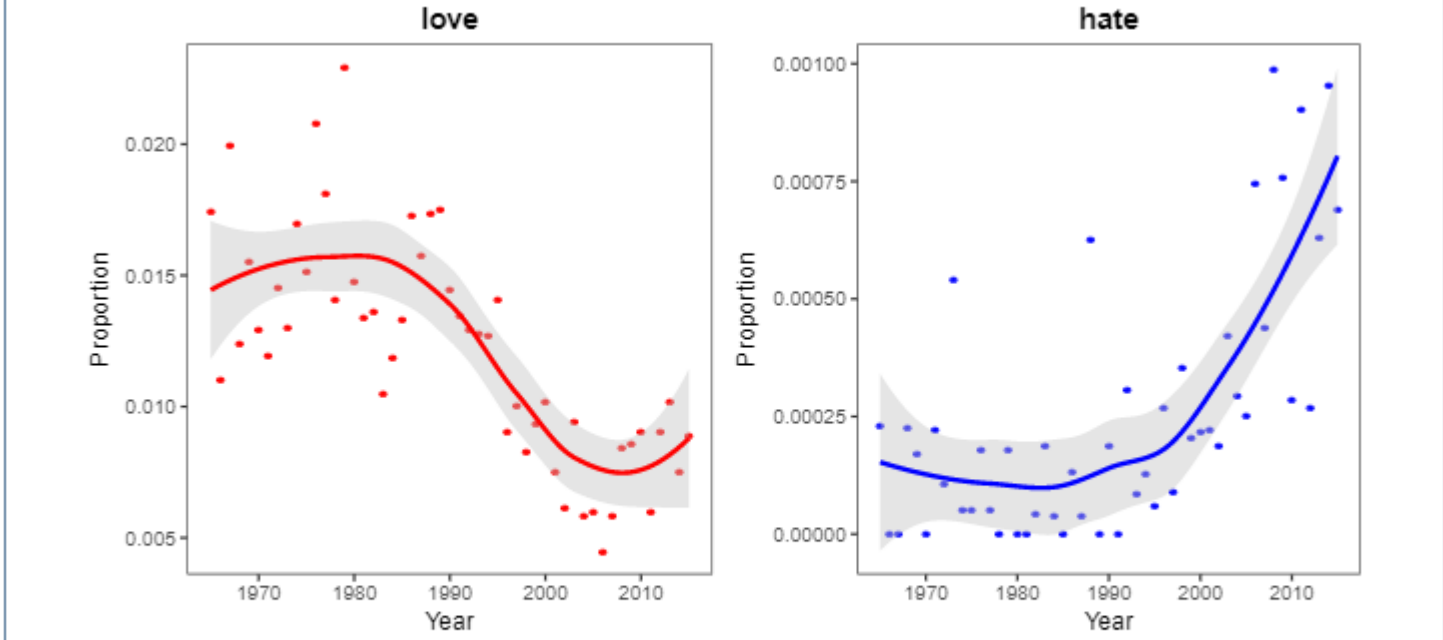


Figure 1 'Love' and 'hate' through years in the Hot 100 Billboard. Proportion of the term 'love' and 'hate' with respect to the year words total, for all the songs in the billboard dataset.

Morin & Acerbi 2016 Birth of the cool

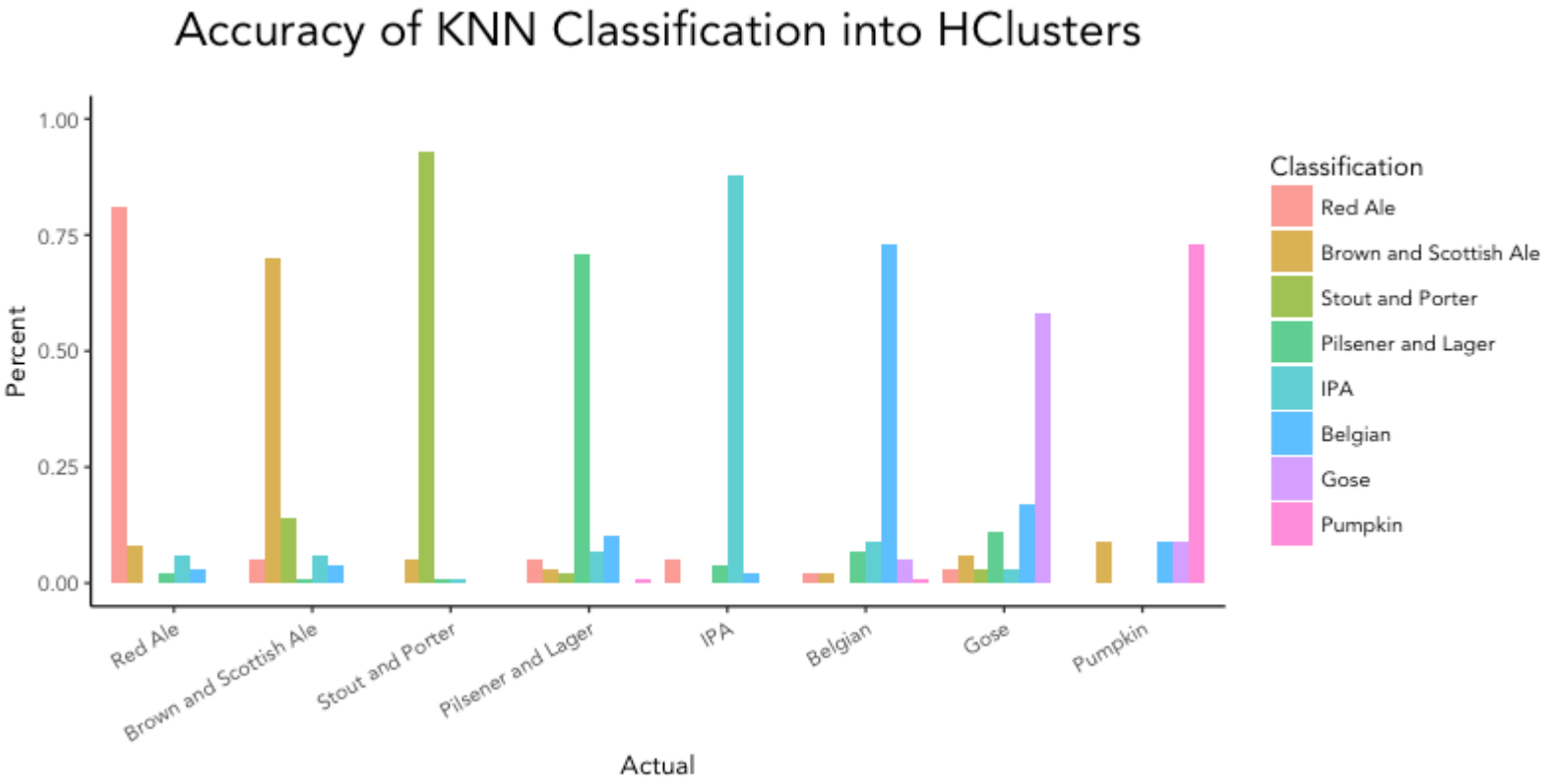
Brand et al. 2018 Cultural evolution of emotional expression in 50 years of song lyrics

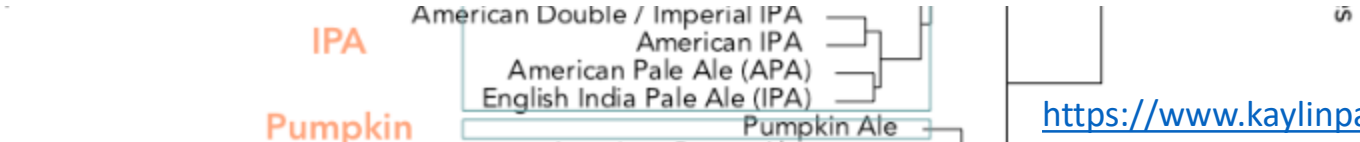
Geo-locations

- Which locations are mentioned in text

Frontini et al. 2016. REDEN ONLINE: Disambiguation, Linking and Visualisation of References in TEI Digital Editions

Word distributions

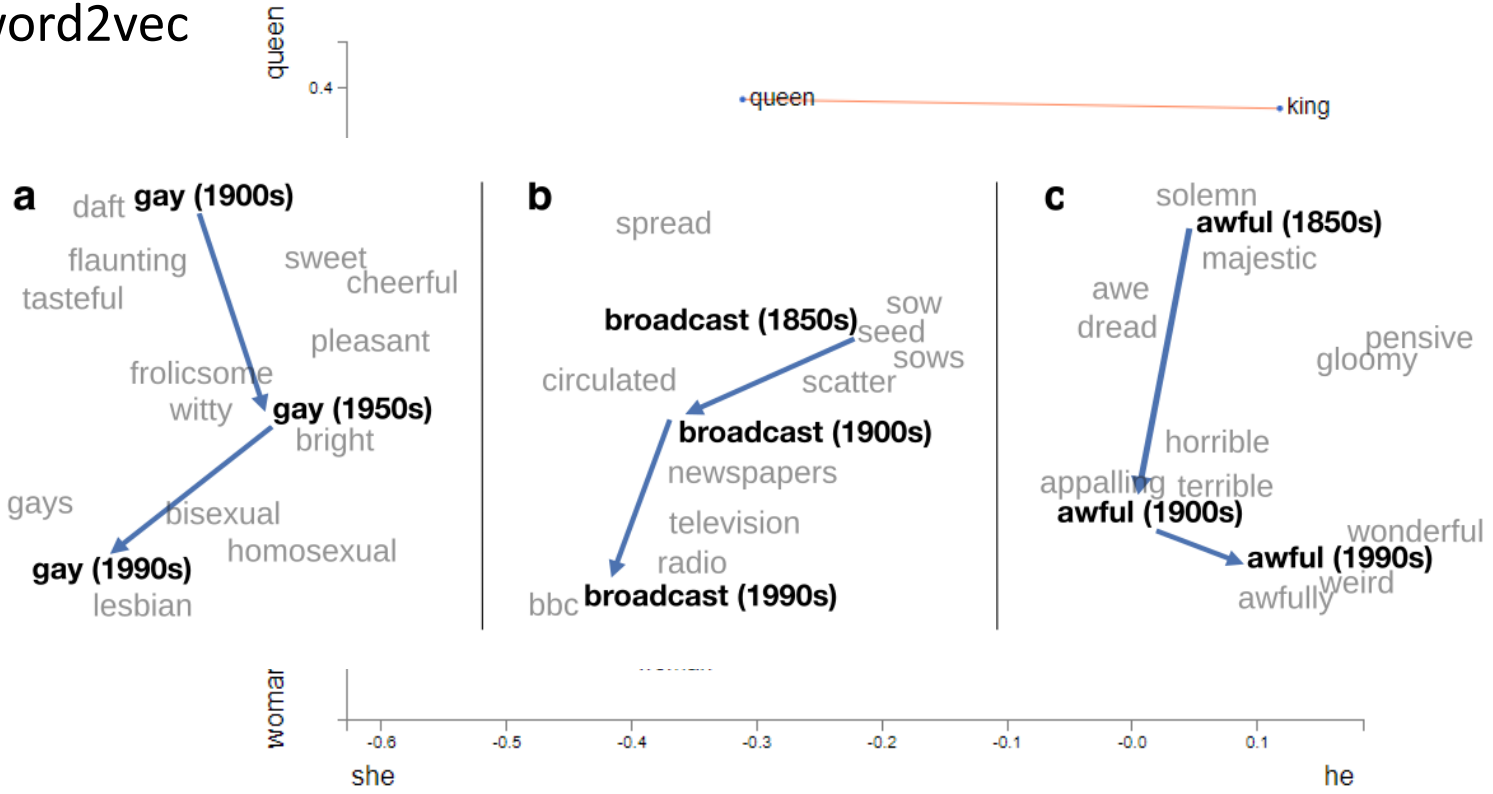




https://www.kaylinpavlik.com/tidy-text-beer/

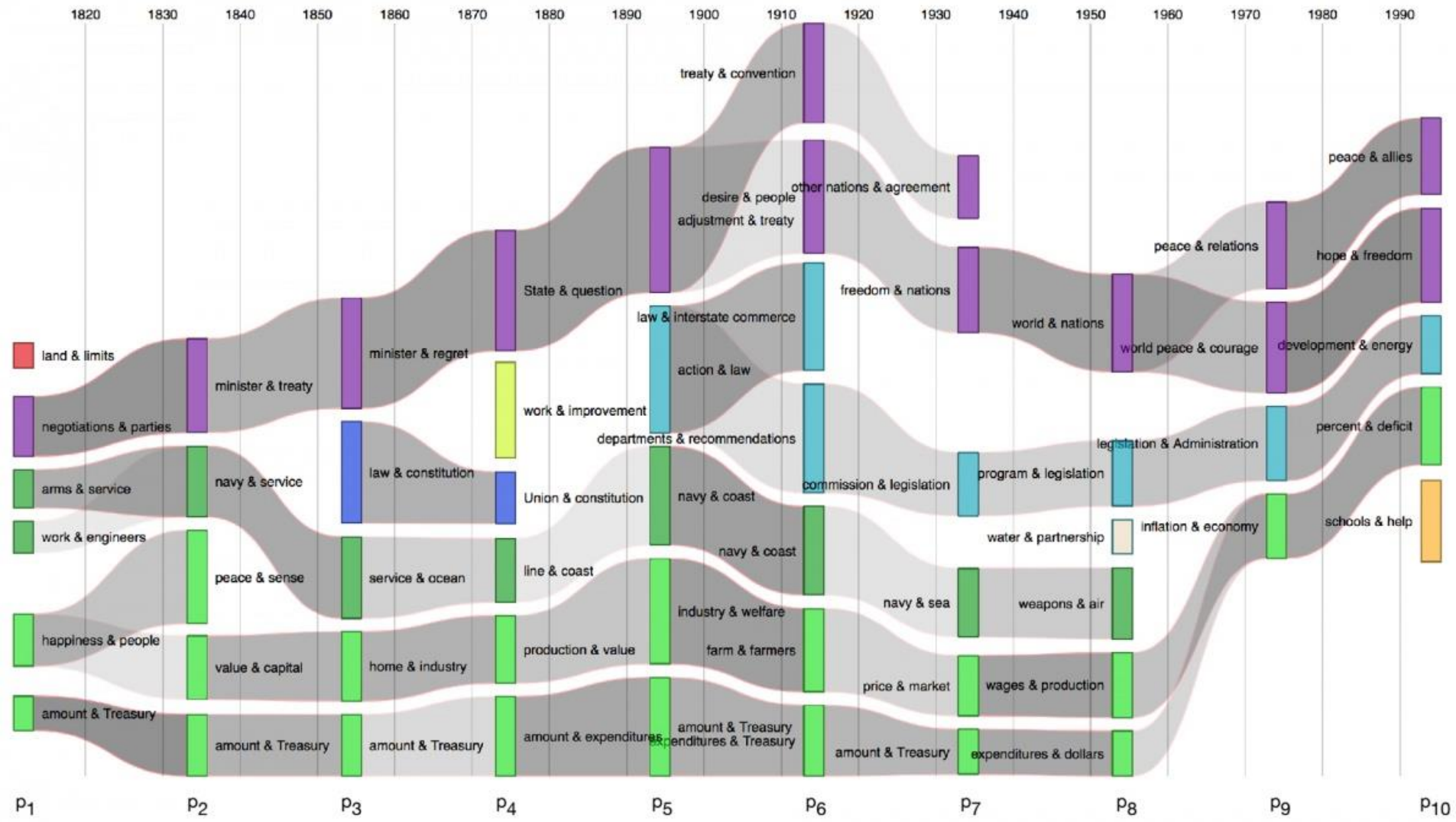
Word semantics

- word2vec



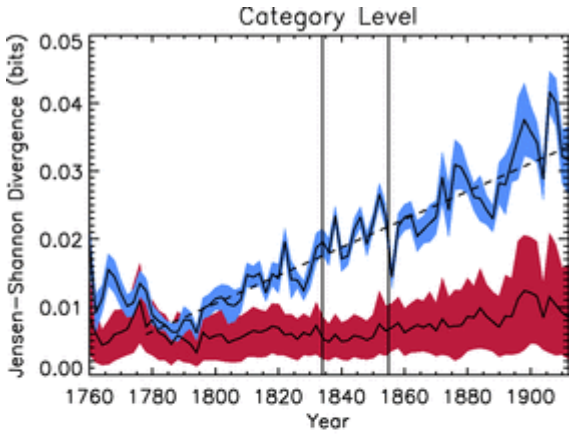
https://lamyiowce.github.io/word2viz/

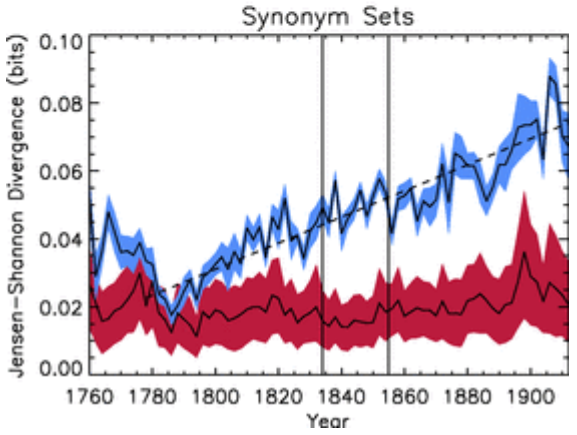
Presidential speechs in USA



Rule et al. 2015 Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014

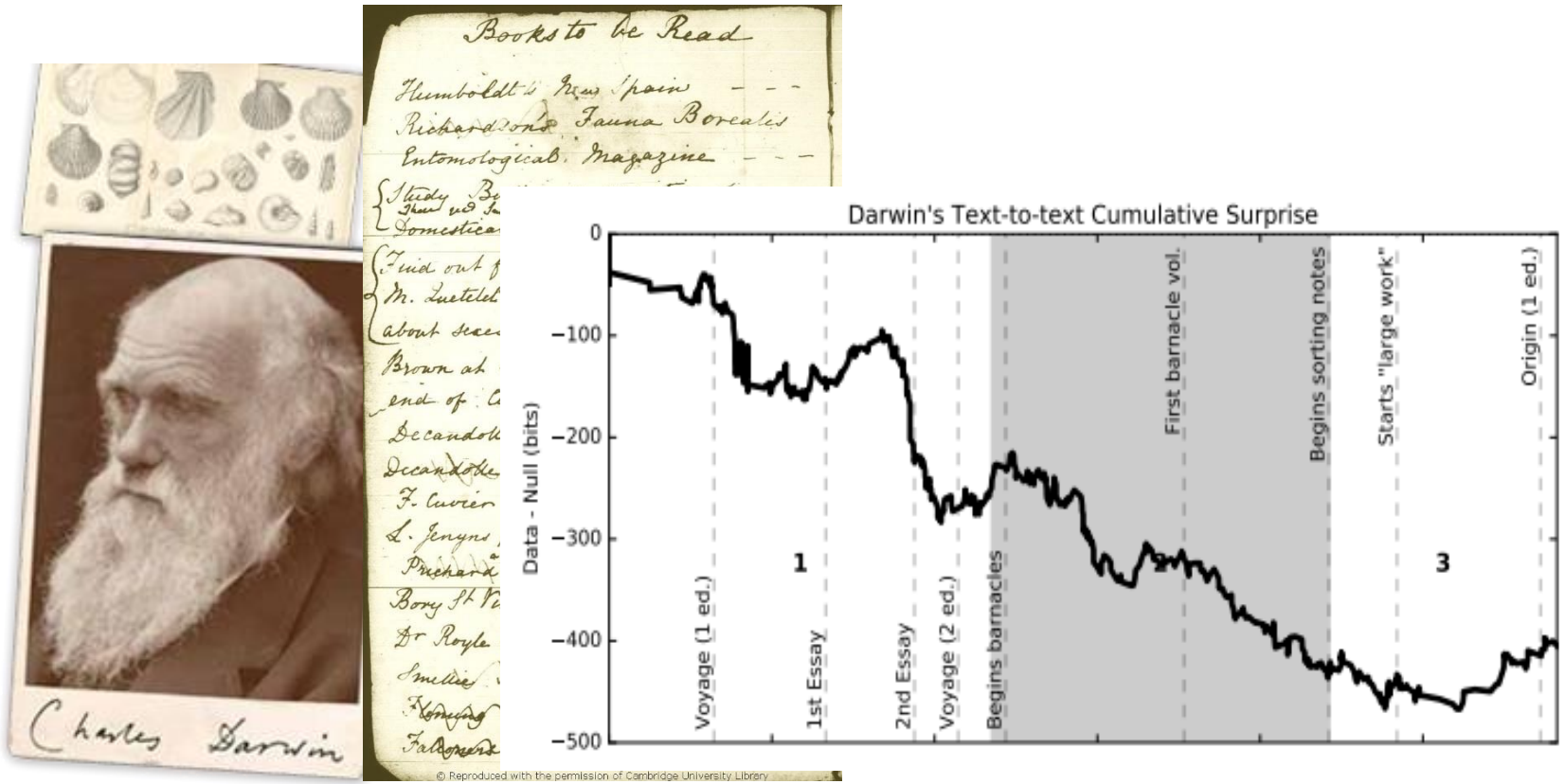
Civilizing process





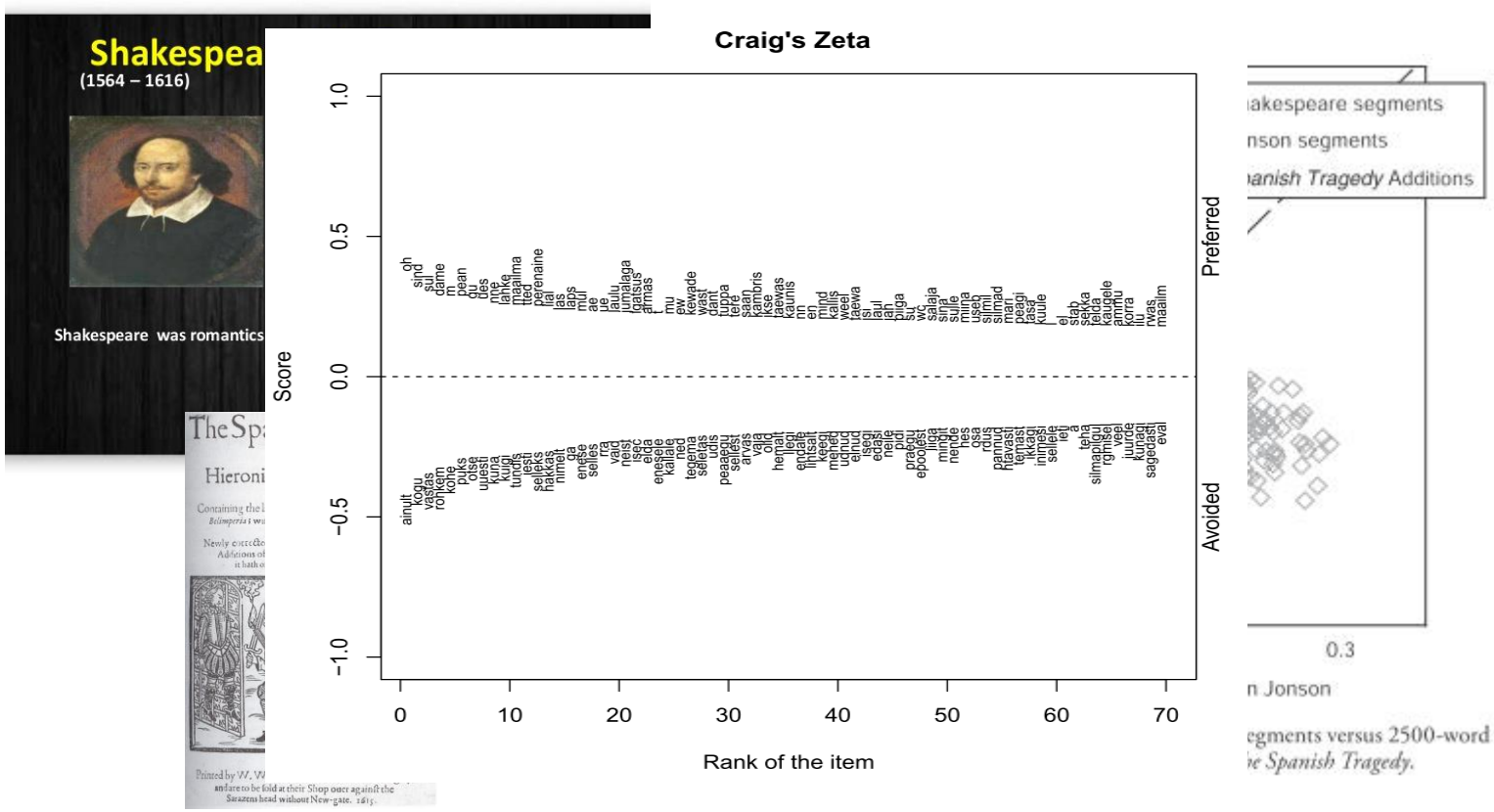
Klingenstein et al. 2014. The civilizing process in London's Old Bailey

Darwin's reading habits



Murdock et al. 2017. Exploration and exploitation of Victorian science in Darwin's reading notebooks.

Finding the author

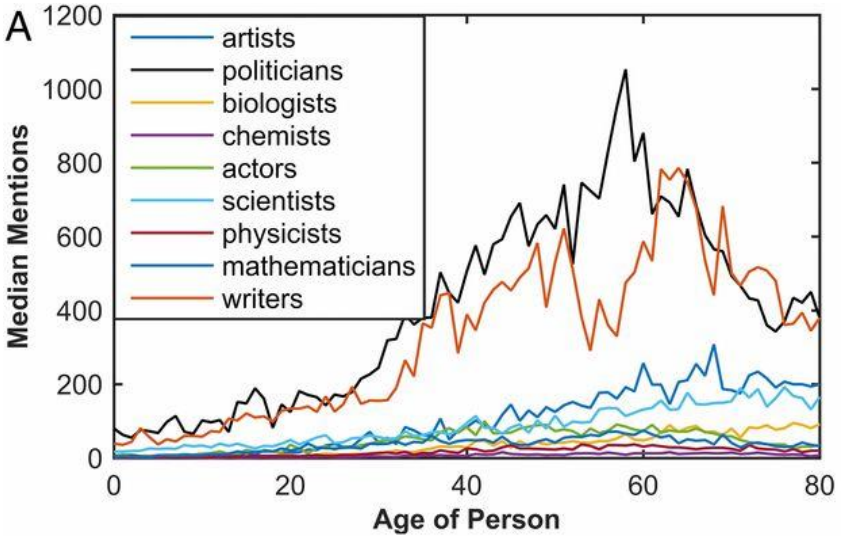


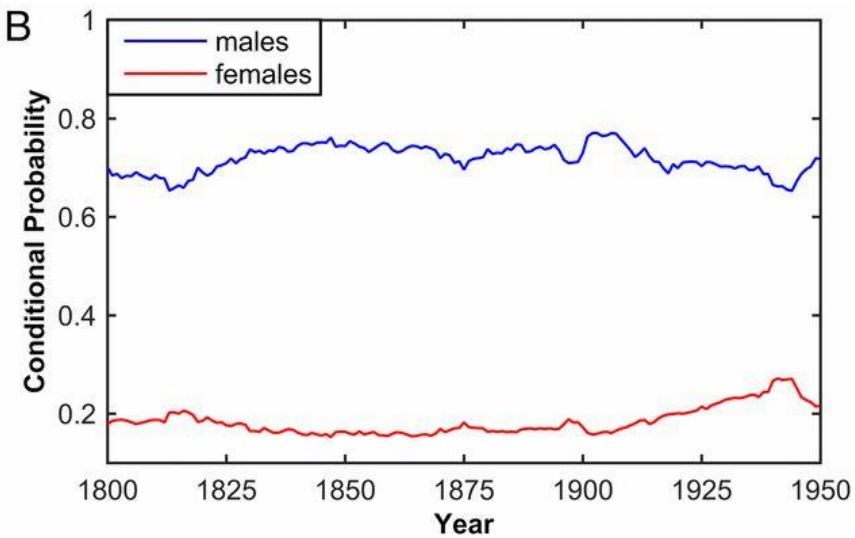
Craig & Kinney 2009. Shakespeare, Computers, and the Mystery of Authorship

Metainformation!

150 years of newspapers

- Linked to biographical data



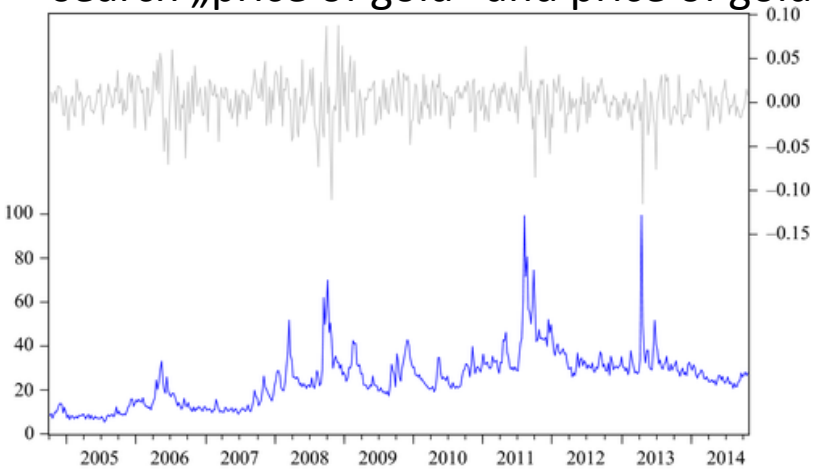


Lansdall-Welfare 2017. Content analysis of 150 years of British periodicals

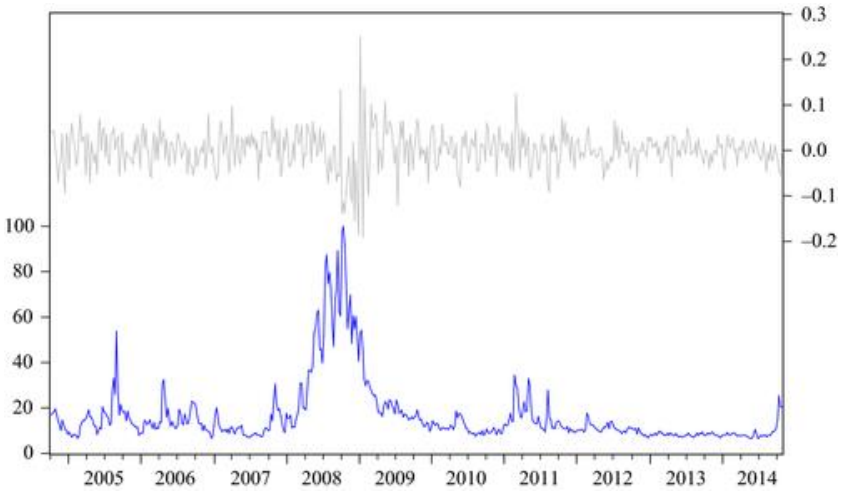
Google searches

- Linked to exact time and economy

search „price of gold“ and price of gold



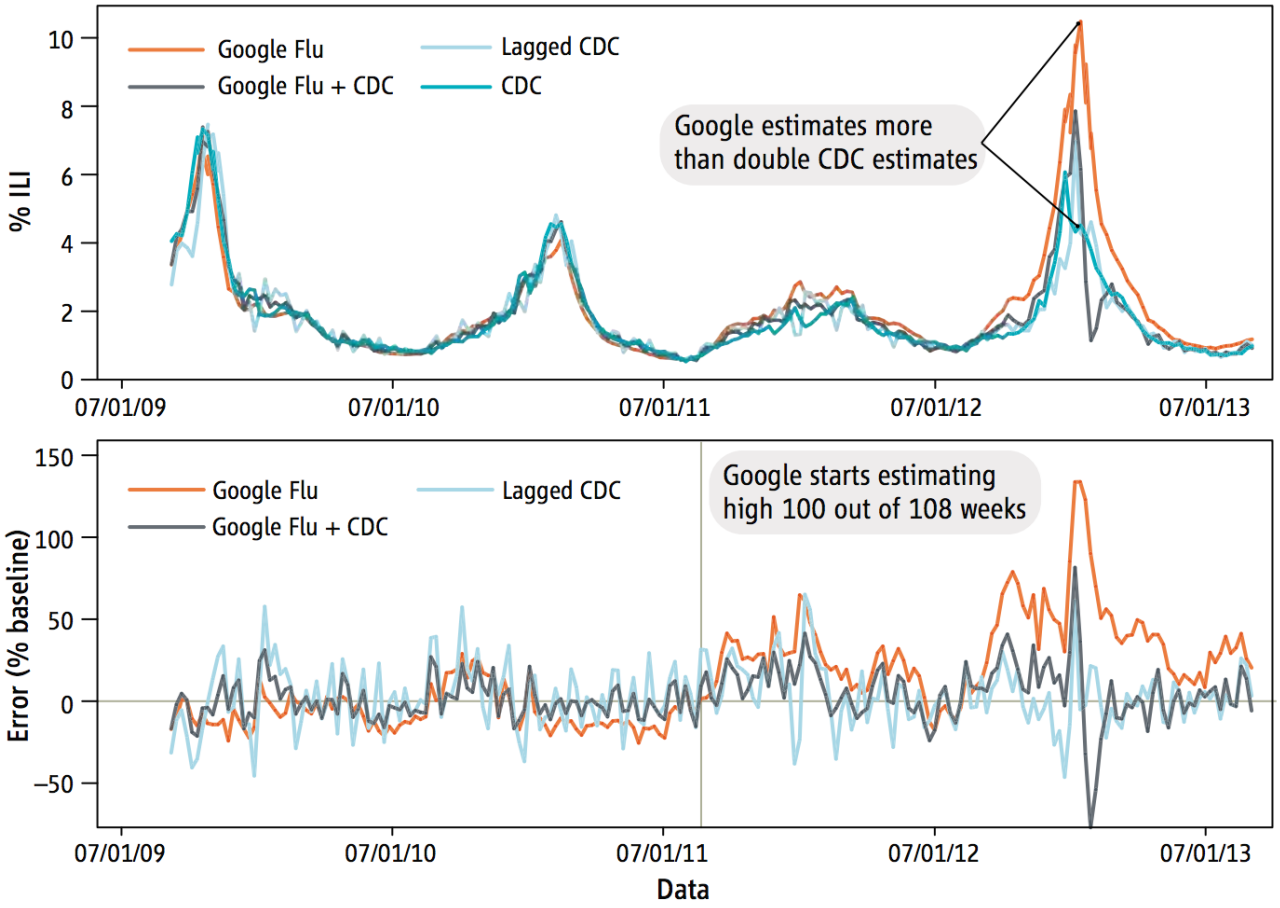
search „price of oil“ and price of oil



Bampinas et al. 2019 Volatility persistence and asymmetry under the microscope: the role of information demand for gold and oil

Google Flu trends

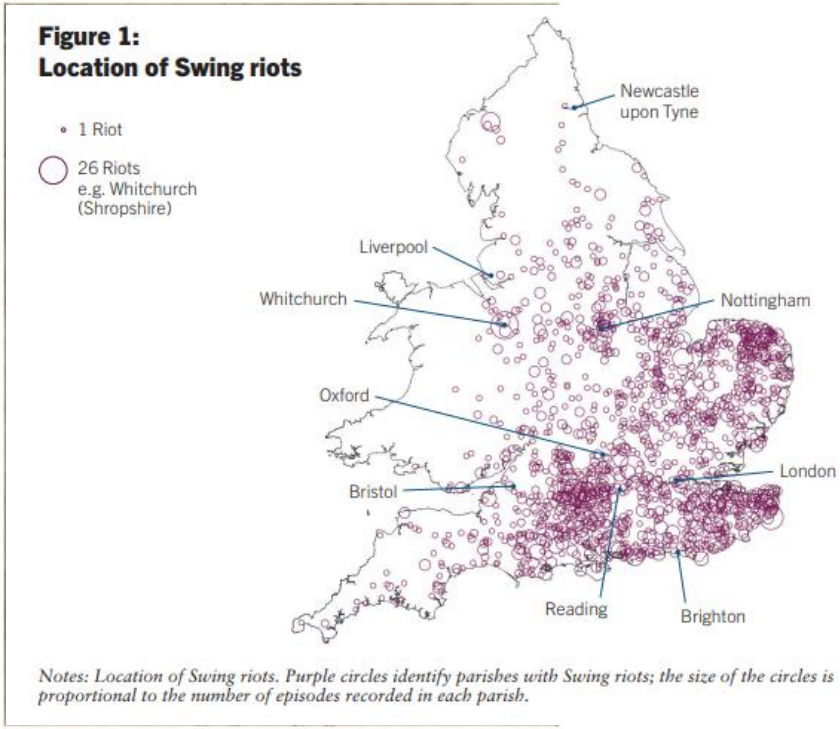
- Search for „flu“ linked with medical estimates

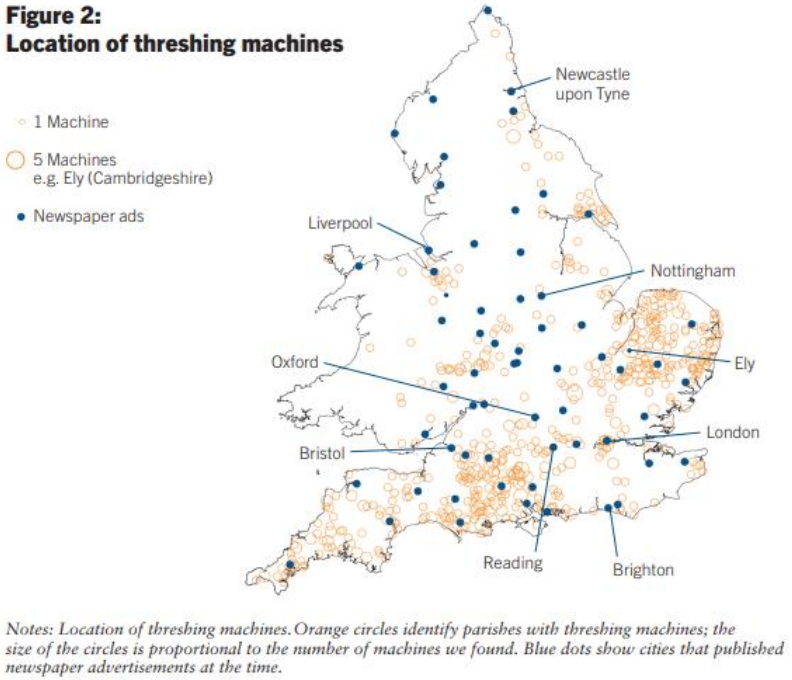


Lazer 2014. The Parable of Google Flu: Traps in Big Data Analysis

Deep mining history

- Advertisements in newspapers -> diffusion of technology
- „Swing riots“ 1829-1830: machines, unemployment and crops.





Caprettini & Voth 2018 Rage against the machines: New technology and violent unrest in industrializing England