Computational text analysis for historians

What? How? (and a tiny bit why?)

NTNU, Trondheim, 12.-13. August 2021

Gregory Ferguson-Cradler
Institutt for rettsvitenskap, filosofi og internasjonale studier
Høgskolen i Innlandet, Lillehammer

What is out there?

Word counts and dictionaries[1]

- Just counting words! (First project in digital humanities (Graham, Milligan, and Weingart, forthcoming))
- The well-known Google n-gram
- Dictionaries for scoring texts (also used for text classification)

1. The following overview follows a longer, more detailed survey of use of computational text analysis methods in a forthcoming paper (Ferguson-Cradler, forthcoming). I've included a pre-print in the workshop Github repository.

Word frequency and dictionary methods in practice

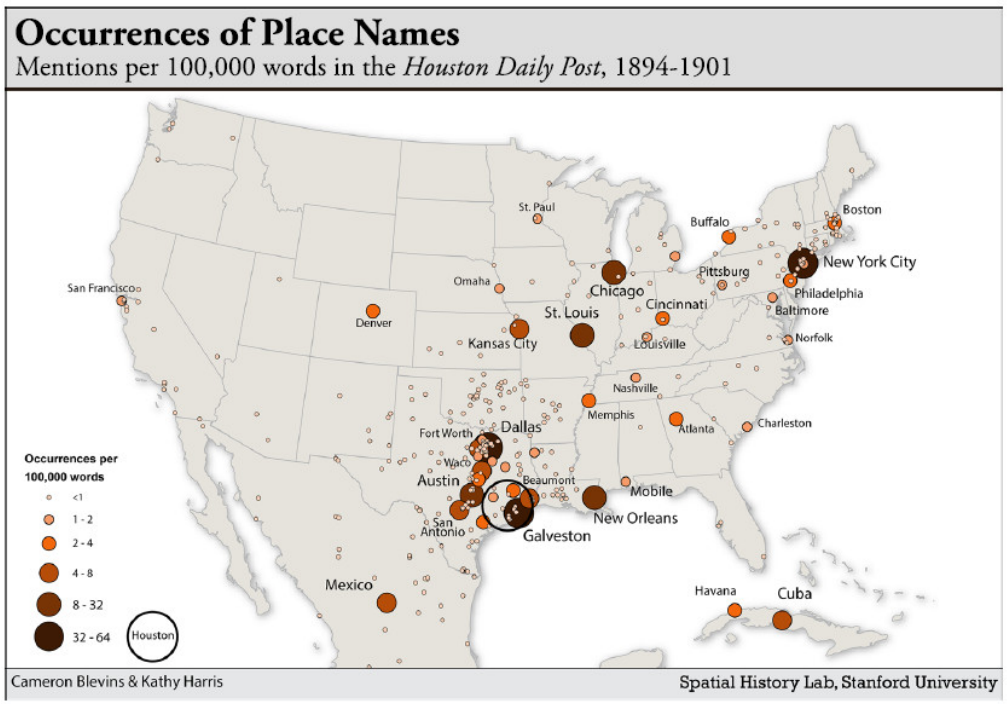


Figure 1. The frequency with which the *Houston Daily Post* printed specific place-names reveals the imagined geography of the newspaper. *Map by Cameron Blevins and Kathy Harris, Spatial History Lab, Stanford University.*

- Tracking news attention in 19th-century America (Blevins 2014)
- Randomize newspaper articles
- Hand-counting
- Mapping of American media attention

Word frequency and dictionary methods in practice, II

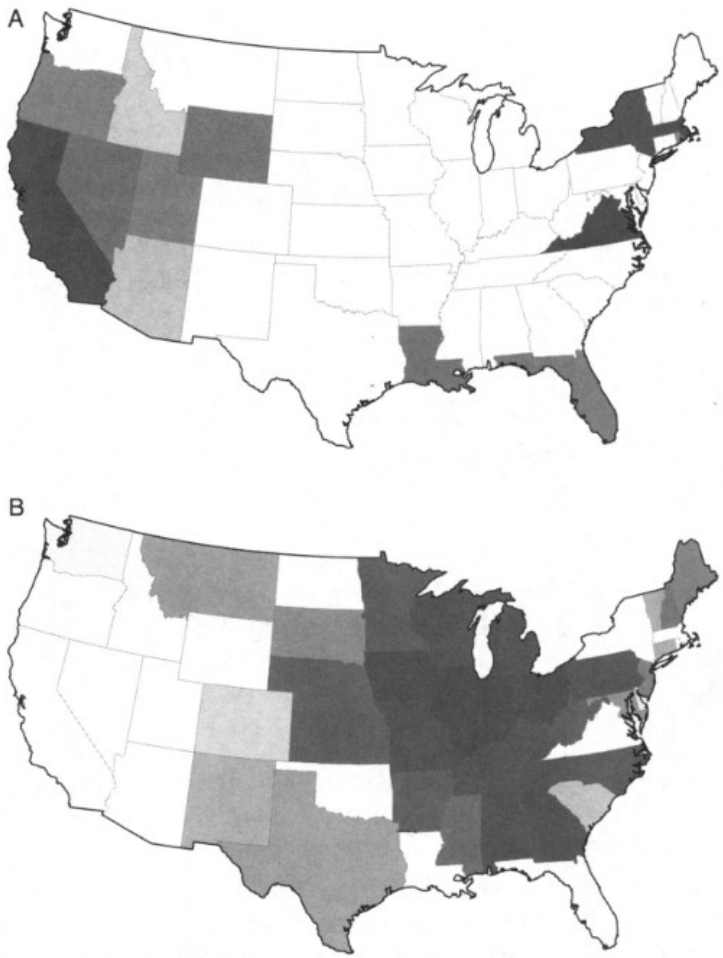
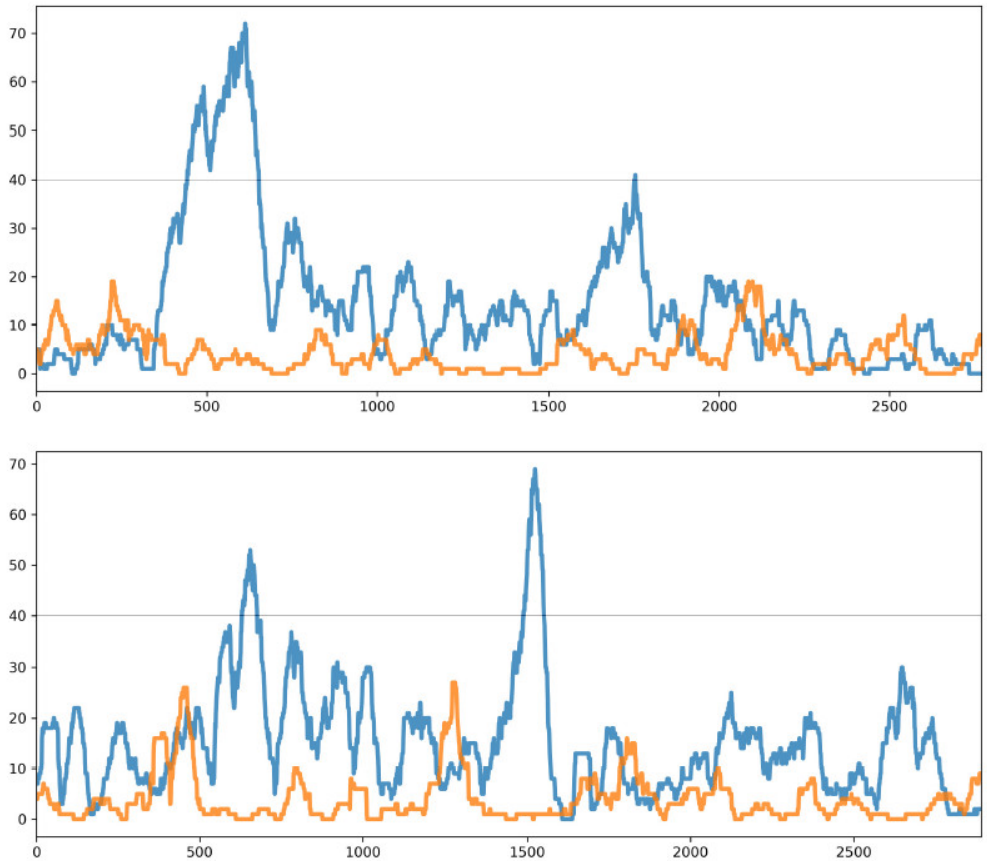


Fig. 7. Dunning log-likelihood values for named-location counts in the full corpus measured against mean state populations, 1850–80. (a) States overrepresented relative to their populations; (b) underrepresented states. Darker shades indicate larger absolute values, hence greater under- or overrepresentation.

- Geographical imagination in American 19th-century fiction (Wilkens 2013)

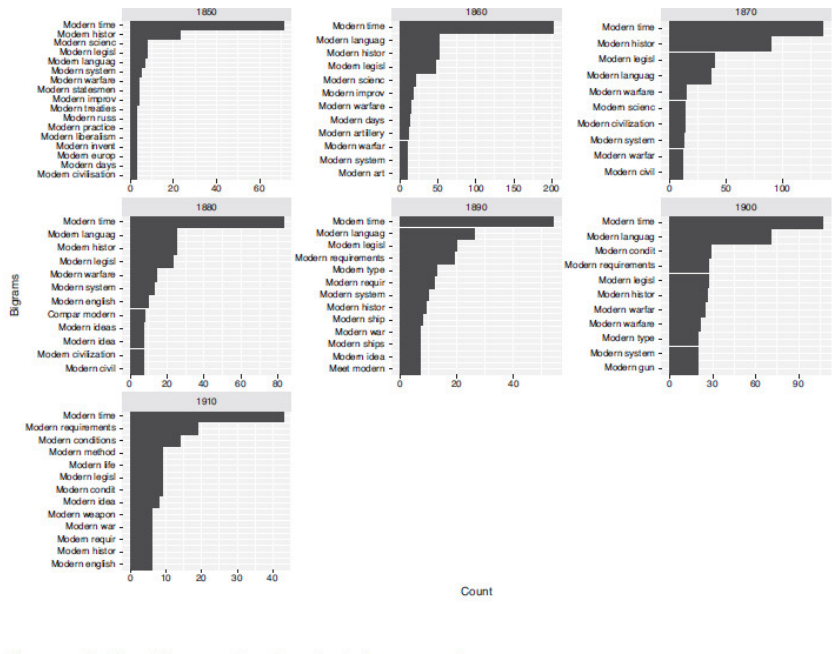
Word frequency and dictionary methods in practice, III



Figur 8–9 To verker som viser narrative mønstre som er forskjellige fra bind to av Randi Rønning Balsviks Vardøverk. I begge tilfeller preges framstillingen av ord assosiert med den offentlige sfæren (blå graf), mens den private sfæren er noe mindre representert (oransje graf) Øverst: Anders Bjarne Fossen *Det nye Askøy blir til 1870–1940*, Askøys historie bind II fra 1996. Nederst: Halvard Tjelmeland: *Fra byfolk og bona til tromsøværing 1945-1996. Tromsø gjennom 10 000 år*. For en nærmere kommentar, se brødteksten.

- Tracking attention to "public" and "private spheres" in Norwegian regional historiography using dictionary word counts (Alsvik and Munch-Møller 2020)

Word frequency and dictionary methods in practice, IV



- Shifting notions of "modernity" via bigram frequency counts (Guldi 2019b)

Document similarity

- Tracing how similar documents are to each other
- Multiple ways to do this: counting text chunks that are exactly the same, charting document similarity over all vocabulary

Document similarity methods in practice

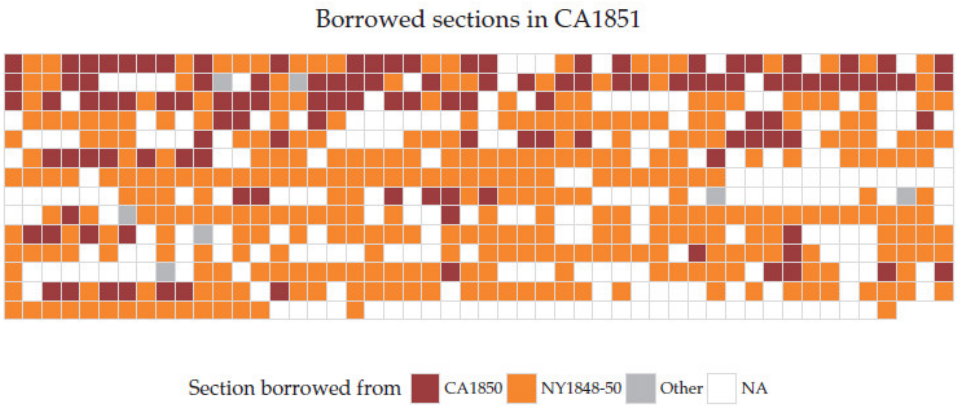


FIGURE 7: When California revised its code of civil procedure in 1851, it borrowed primarily from the 1850 New York code and not its own earlier 1850 code.

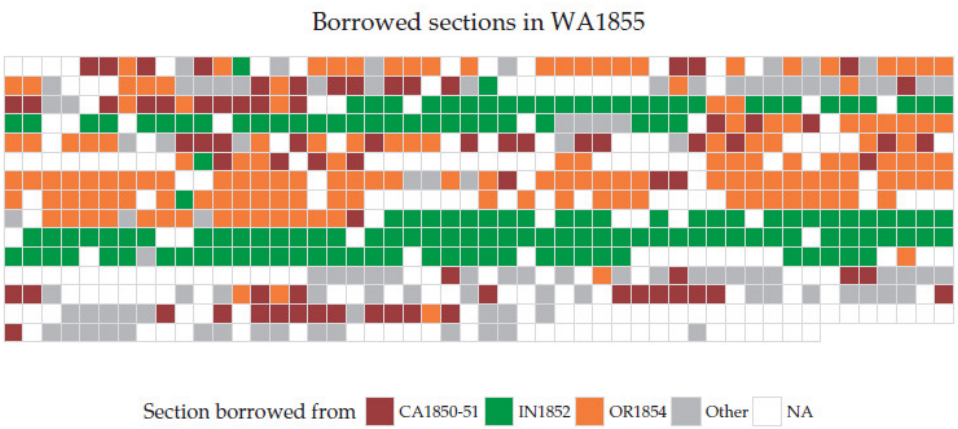
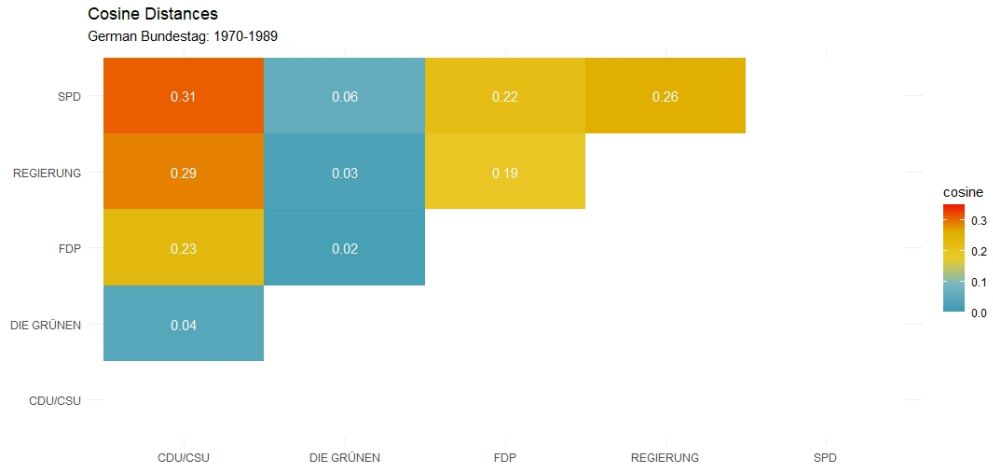
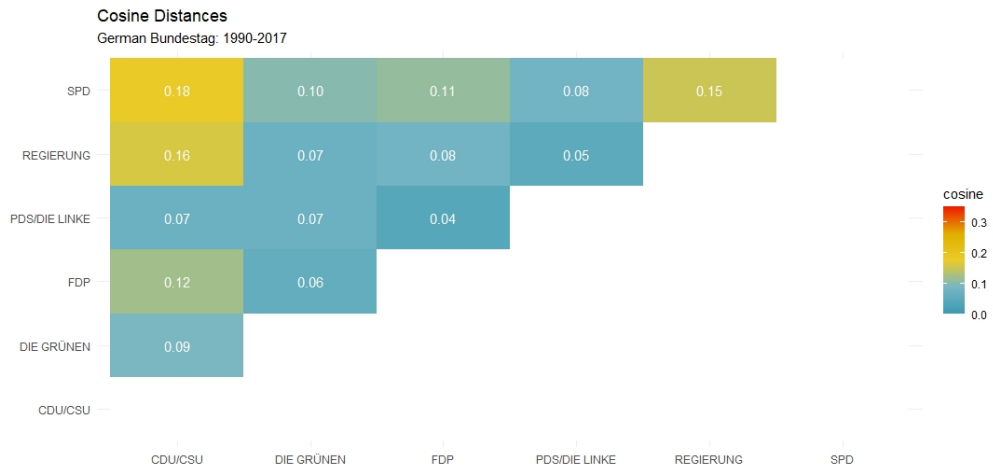


FIGURE 8: Washington's 1855 code of civil procedure borrowed long contiguous sections from Indiana's 1852 code and Oregon's 1854 code. Washington's code commissioners had previously been judges in those jurisdictions, which also borrowed their procedure codes from New York's Field Code.

- Civil code adoption that can be uncovered by looking for identical sections of civil codes in 19th-century United States (Funk and Mullen 2018)

Document similarity methods in practice, II





- Tracing vocabulary similarity in vector space of statements in the German Bundestag by party (Ferguson-Cradler, forthcoming)

Document similarity

- Tracing themes/topics through documents
- Each text is seen as a mixture of topics, and each topic as a mixture of words
- Long been the most visible and well-known method in digital humanities

Topic models in practice

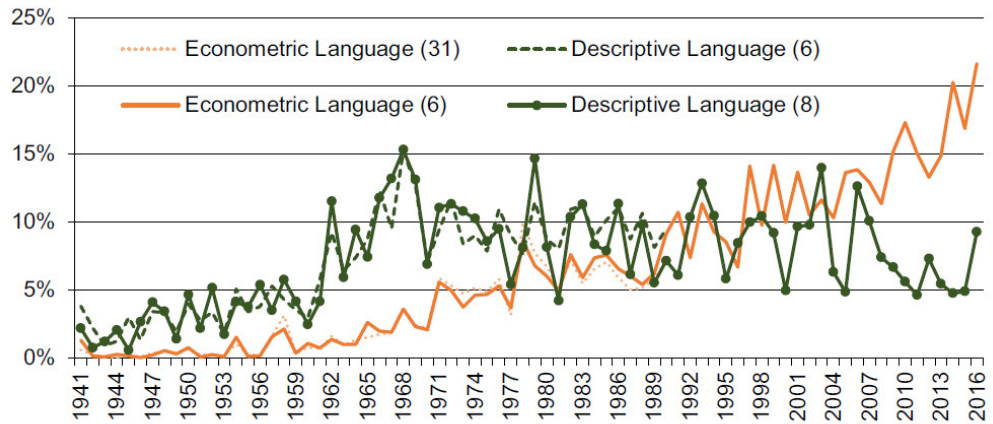


Fig. 6 Topic shares of quantitative topics. Dotted lines mark topics from sample 1 and solid lines mark topics from sample 2; annual means. *Source*: See text

- Topic modeling the *Journal of Economic History* to find shift in language to quantification as a topic (Wehrheim 2019)

Topic models in practice, II

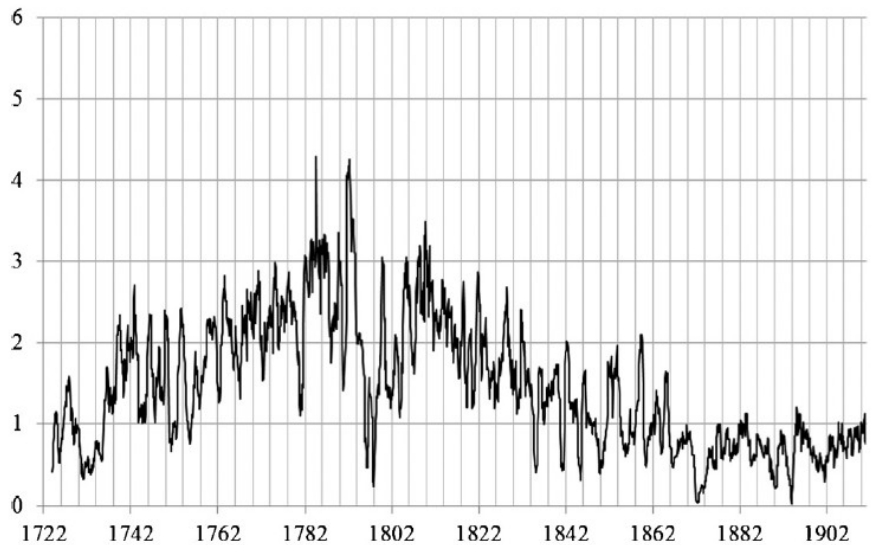


Fig. 1. Crime topic proportion over time (12 month moving average of monthly sum).

- Topic models to reconstruct a "crime rate" for 19th-century China (Miller 2013)

Topic models in practice, III

TABLE 2

SELECTED INFRASTRUCTURE TOPICS FROM A 500-TOPIC MODEL OF HANSARD, 1800–1910

Topic number	*Probability*	*Words in order of prominence*
		HIGHWAY REGULATION TOPIC
39	0.00547	road-, roads-, public-, mile-, highway-, motor-, carriage-, toll-, turnpike-, speed-, traffic-, car-, horse-, repair-, cab-, main-, driver-, trust-, limit-, vehicle-
		POST OFFICE ADMINISTRATION TOPIC
164	0.00467	mail-, service-, post-, general-, service-, postmast-, postal-, letter-, mails-, train-, contract-, arrange-, convey-, London-, company-, office-, packet-, railway-, delivery-, arrive-
		RIVER INFRASTRUCTURE TOPIC
190	0.00466	river-, drainage-, water-, work-, drainage-, board-, navig-, sewag-, shannon-, thame-, conserve-, navigation-, thames-, canal-, works-, flood-, carri-, land-, district-, improv-
		RAILWAY TOPIC
4	0.00936	railway-, line-, company-, railways-, construct-, great-, light-, western-, interest-, company-, public-, mile-, scheme-, traffic-, guarante-, work-, railroad-, companies-, promot-, propos-

- What was the British state talking about when it was talking about infrastructure? (Guldi 2019a)

Word embedding

- Represents words as vectors in many-dimensional vector space
- Dimension reduction can be used to plot words in relation to each other (over time or space)
- Axes that correspond with meaning can be constructed and words placed on this spectrum

Word embedding in practice

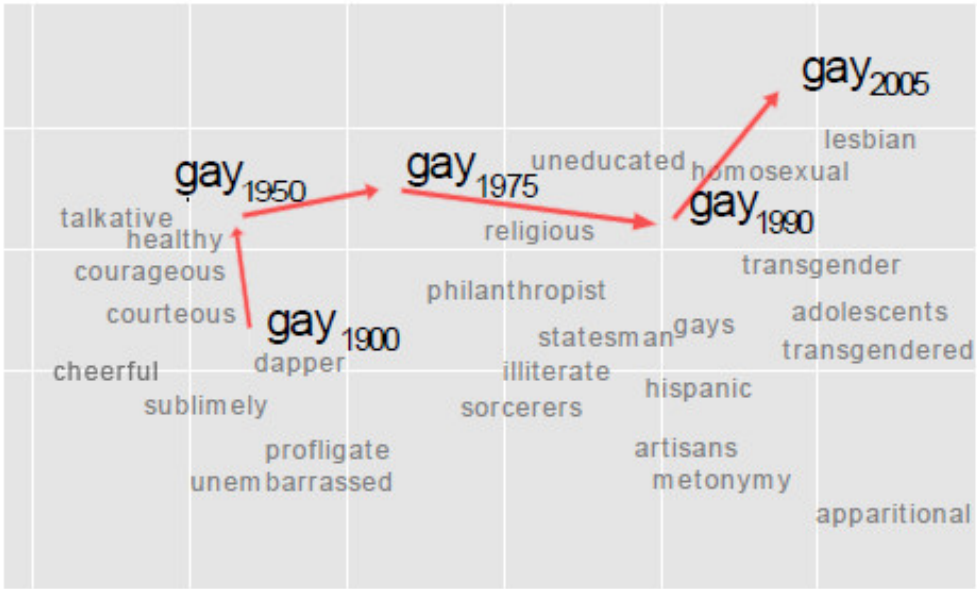


Figure 1: A 2-dimensional projection of the latent semantic space captured by our algorithm. Notice the semantic trajectory of the word **gay** transitioning meaning in the space.

- Visualizing the changing meaning of words over time (Kulkarni et al. 2015)

Workshop goals

- How to read documents into R
- How to make analyze and visualize word frequency
- Use dictionaries to track word use, do sentiment categorization
- Do basic document similarity analysis and topic modeling

Workshop schedule: Thursday, August 12

9:00 – 9:30	Introductions and workshop goals
9:30 – 10:30	Session 1: Basics of R
10:30 – 10:45	Break
10:45 – 12:15	Session 2: Web scraping and reading documents into R
12:15 – 1:15	Lunch
1:15 – 3:00	Session 3: Cleaning and manipulating text in R
3:00 – 3:30	Break
3:30 – 5:00	Session 4: Word frequencies, word clouds, and basic counting
5:00 –	Troubleshooting

Workshop schedule: Friday, August 13

9:00 – 10:30	Session 5: POS and dictionary methods
10:30 – 10:45	Break
10:45 – 12:15	Session 6: Document similarity and co-occurrence
12:15 – 1:15	Lunch
1:15 – 3:00	Session 7: Topic models: theory and practice
3:00 – 3:30	Break
3:30 – 4:30	Continuation of Session 7: Topic models
4: 30 – 5:15	Conclusion: Discussion of other techniques of possible interest and conclusion

Alsvik, Ola, and Marthe Glad Munch-Møller. 2020. Historiografi møter algoritmer. *Heimen* 57 (3): 201–215.

Blevins, Cameron. 2014. Space, nation, and the triumph of region: a view of the world from houston. *The Journal of American History* 101 (1): 122–147.

Ferguson-Cradler, Gregory. Forthcoming. Narrative and computational text analysis in business and economic history. *Scandinavian Review of Economic History.*

Funk, Kellen, and Lincoln A. Mullen. 2018. The Spine of American Law: Digital Text Analysis and U.S. Legal Practice. *The American Historical Review* 123 (1): 132–164.

Graham, Shawn, Ian Milligan, and Scott Weingart. Forthcoming. *Exploring big historical data: the historian's macroscope.* 2nd. World Scientific Publishing Company.

Guldi, Jo. 2019a. Parliament's debates about infrastructure: an exercise in using dynamic topic models to synthesize historical change. *Technology and culture* 60 (1): 1–33.

———. 2019b. The measures of modernity: the new quantitative metrics of historical change over time and their critical interpretation. *International Journal for History, Culture and Modernity* 7 (1): 899–939.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web*, 625–635.

Miller, Ian Matthew. 2013. Rebellion, crime and violence in qing china, 1722–1911: a topic modeling approach. *Poetics* 41 (6): 626–649.

Wehrheim, Lino. 2019. Economic history goes digital: topic modeling the journal of economic history. *Cliometrica* 13 (1): 83–125.

Wilkens, Matthew. 2013. The geographic imagination of civil war-era american fiction. *American literary history* 25 (4): 803–840.