Computational text analysis for the humanities

What and how?

NTNU, Trondheim, 9. november 2021

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

Workshop schedule

9:00 – 10:30	Overview of quantitative text methods and uses
10:30 – 10:45	Break
10:45 – 11:50	Brief overview and introduction to R
11:50 - 12:00	Conclusion

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 2021)

Word frequency and dictionary methods in practice

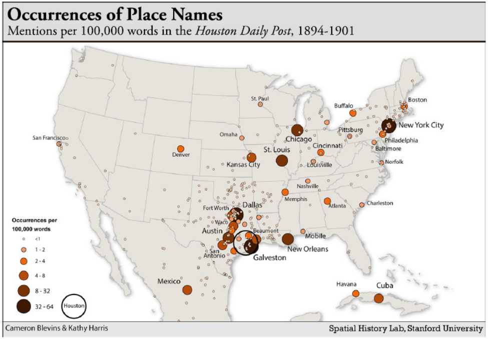


Figure 1. The frequency with which the *Houston Daily Post* printed specific place-names reveals the imagined geography of the newspaper. *Map by Cameron Blevins and Kathy Harris, Spatial History Lab, Stanford University.*

- 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

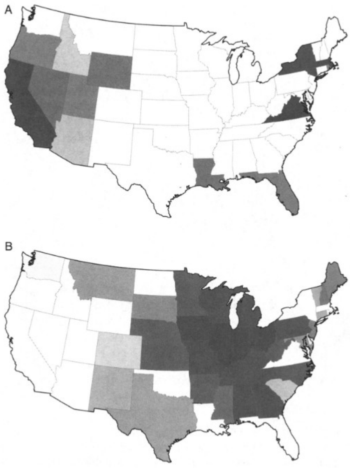
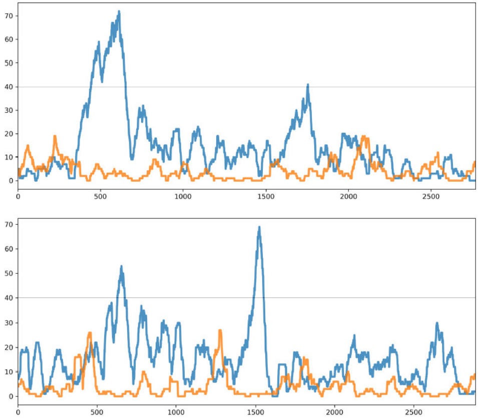


Fig. 7. Dunning log-likelihood values for named-location counts in the full corpus measured against mean state populations, 1850–80. (a) States overrepresented relative to their populations; (b) underrepresented states. Darker shades indicate larger absolute values, hence greater under- or overrepresentation.

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

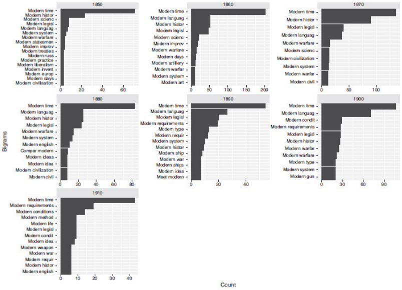
Word frequency and dictionary methods in practice, III



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

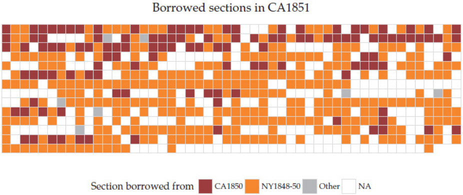


FIGURE 7: When California revised its code of civil procedure in 1851, it borrowed primarily from the 1850 New York code and not its own earlier 1850 code.

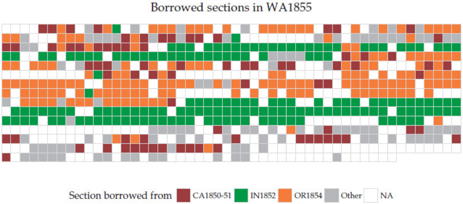


FIGURE 8: Washington's 1855 code of civil procedure borrowed long contiguous sections from Indiana's 1852 code and Oregon's 1854 code. Washington's code commissioners had previously been judges in those jurisdictions, which also borrowed their procedure codes from New York's Field Code.

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

Cosine similarity

Cosine similarity involves treating documents as lists (vectors) of words.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.3 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.

Cosine similarity

Cosine similarity involves treating documents as lists (vectors) of words.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.3 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.

We can (try) to think of such vectors spatially and measure the angle between them.

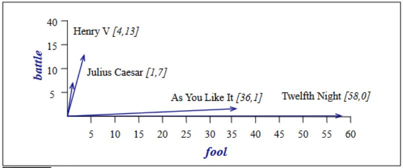
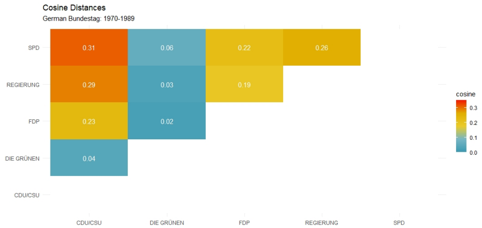
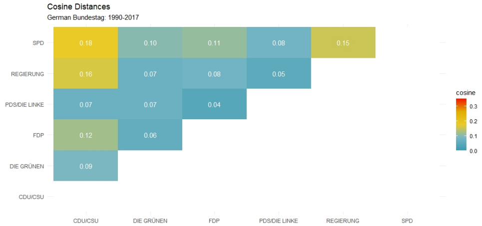


Figure 6.4 A spatial visualization of the document vectors for the four Shakespeare play documents, showing just two of the dimensions, corresponding to the words *battle* and *fool*. The comedies have high values for the *fool* dimension and low values for the *battle* dimension.

(Jurafsky and Martin, forthcoming, ch. 6.3)

Document similarity methods in practice, II



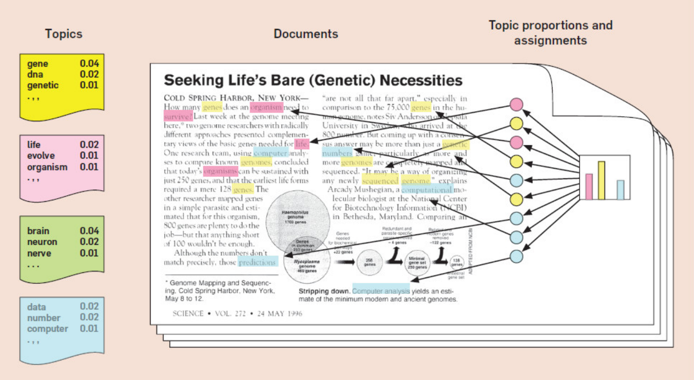


- Tracing vocabulary similarity in vector space of statements in the German Bundestag by party (Ferguson-Cradler 2021)

Topic models

- 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 model algorithm



Text is produced by choosing a distribution of topics within the given document; the for every word a selection of topic based on the document-level distribution; finally a word from the corresponding topic (Blei 2012, 78).

Topic models in practice

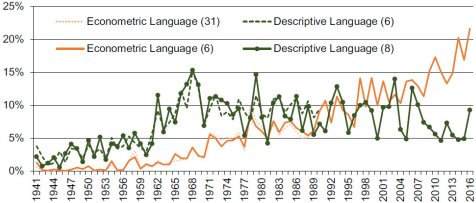


Fig. 6 Topic shares of quantitative topics. Dotted lines mark topics from sample 1 and solid lines mark topics from sample 2; annual means. *Source*: See text

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

Topic models in practice, II

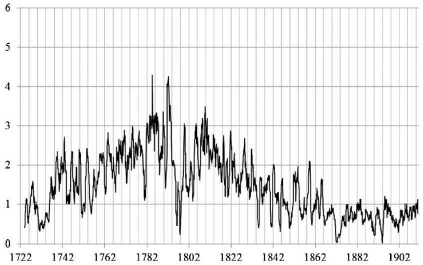


Fig. 1. Crime topic proportion over time (12 month moving average of monthly sum).

- 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 similarity and relatedness

- Based on the ideas that similar words appear in the same context (both words that are synonyms and words that are simply clearly of the same kind, eg. "Germany" and "France").[2]
- Based on the idea that word meaning can be represented in vector space (as we saw in document similarity) based on contexts in which words appear.
- Documents made into vectors via DTM matrix. Words might be made into vectors via term-term matrix (fcm in Quanteda)
- Two major algorithms for word embedding: word2vec and GloVe.

2. This is based on long and deep thought in linguistics, see (Jurafsky and Martin, forthcoming) for a brief overview.

What *are* word embeddings

- Simplifying: these algorithms compute probability for word co-occurrences (and non-co-occurrences) and construct word embeddings (vectors) that are similar when co-occurrence probability is high and distant when probability is low.[3]

3. Jurafsky and Martin (forthcoming) is the best introduction to the details.

What *are* word embeddings

- Simplifying: these algorithms compute probability for word co-occurrences (and non-co-occurrences) and construct word embeddings (vectors) that are similar when co-occurrence probability is high and distant when probability is low.[3]
- Word embeddings so interesting (and somewhat baffling) because they show not just similarities between words but also have vector spaces that seem to correspond to meaningful concepts.

3. Jurafsky and Martin (forthcoming) is the best introduction to the details.

What *are* word embeddings

- Simplifying: these algorithms compute probability for word co-occurrences (and non-co-occurrences) and construct word embeddings (vectors) that are similar when co-occurrence probability is high and distant when probability is low.[3]
- Word embeddings so interesting (and somewhat baffling) because they show not just similarities between words but also have vector spaces that seem to correspond to meaningful concepts.
- $\overrightarrow{king} + \overrightarrow{woman} - \overrightarrow{man} \approx \overrightarrow{queen}$ analogous to just as a human would generally suggest 'queen' in answer to the question: man:woman as king:______?.

3. Jurafsky and Martin (forthcoming) is the best introduction to the details.

Word embedding in practice, I

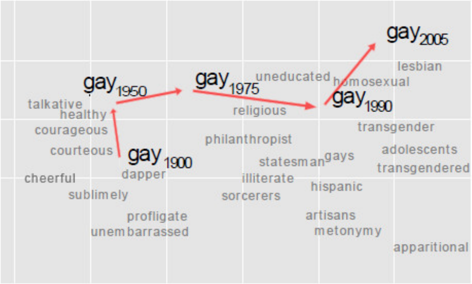


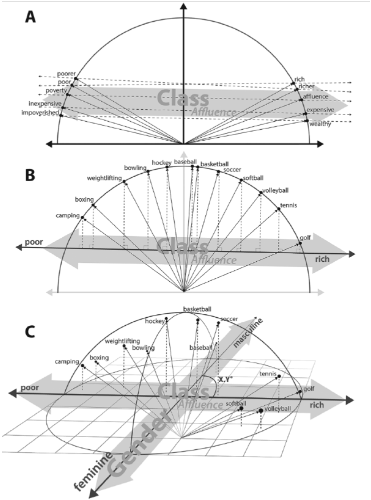
Figure 1: A 2-dimensional projection of the latent semantic space captured by our algorithm. Notice the semantic trajectory of the word **gay** transitioning meaning in the space.

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

Word embedding in practice, II

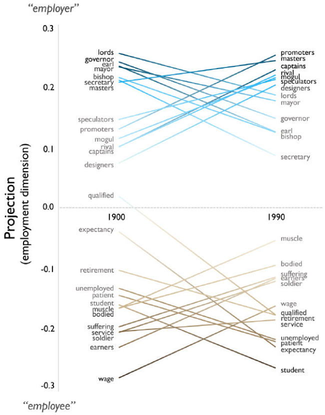
- Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: analyzing the meanings of class through word embeddings. *American Sociological Review* 84 (5): 905–949
- Insight: we find dimensions in vector space that map to human meanings (eg, affluence, etc) by taking the average of pairs of words whose meanings diverge on this range (for affluence: affluence-poverty; rich-poor, prosperous-bankrupt, etc).
- Other words can then be "projected" along this dimension to measure where they stand on the spectrum.

Kozlowski et al. 2019



Kozlowski, Taddy, and Evans 2019

Word embedding in practice, III



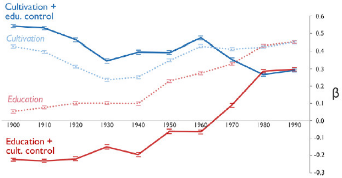


Figure 6. Standardized Coefficients from OLS Regression Models in Which Word Projections on Cultivation and Education Dimensions Predict Projection on the Affluence Dimension; 1900 to 1999 Google Ngrams Corpus
Note: A separate OLS regression model is fit for each decade; N = 50,000 most common words in each decade.

(Kozlowski, Taddy, and Evans 2019, 928, 924)

Contextualized word emebeddings and transformers

- Individual *token* embeddings
- Transformers weigh context word affect words of interest
- So far main use in social science seen for classification
- Unclear how might be taken up for interpretive purposes

Break!

Further resources: textbooks on R and text analysis

- Graham, Milligan, and Weingart, forthcoming
- Learning basics of R: https://swirlstats.com/students.html
- Hadley Wickham and Garrett Grolemund. 2016. *R for data science: import, tidy, transform, visualize, and model data.* O'Reilly Media, Inc. https://r4ds.had.co.nz/
- Julia Silge and David Robinson. 2017. *Text mining with R: a tidy approach.* O'Reilly Media, Inc. https://www.tidytextmining.com/
- Jurafsky and Martin, forthcoming
- Matthew L Jockers and Rosamond Thalken. 2020. *Text analysis with R.* Springer

Further resources: online courses in programming and R

- Introduction to Computer Science and Programming: solid introduction to basics of programming (in Python but easily applicable to R)
- Data analysis for social scientists: basic quantitative methods in social science in R.
- Intro to Data Science: very basic course in R and data science.

Further resources: people in digital humanities/social science to follow

- Ben Schmidt
- Ted Underwood
- Julia Silge
- David Robinson
- Ken Benoit

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

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* 55 (4): 77–84.

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. 2021. 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. http://www.themacroscope.org/2.0/.

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.

Jockers, Matthew L, and Rosamond Thalken. 2020. *Text analysis with R.* Springer.

Jurafsky, Dan, and James H Martin. Forthcoming. *Speech and language processing. vol. 3.* https://web.stanford.edu/~jurafsky/slp3/.

Kozlowski, Austin C, Matt Taddy, and James A Evans. 2019. The geometry of culture: analyzing the meanings of class through word embeddings. *American Sociological Review* 84 (5): 905–949.

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.

Silge, Julia, and David Robinson. 2017. *Text mining with R: a tidy approach.* O'Reilly Media, Inc. https://www.tidytextmining.com/.

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

Wickham, Hadley, and Garrett Grolemund. 2016. *R for data science: import, tidy, transform, visualize, and model data.* O'Reilly Media, Inc. https://r4ds.had.co.nz/.

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