Text (as Data)

William Lowe

Hertie School

7th October 2020

Why text?

Applications:

→ Agenda measurement e.g. Grimmer et al., 2011
→ Framing studies e.g. Gamson and Modigliani, 1989
→ Authorship attribution e.g. Mosteller and Wallace, 1963
→ Bias measurement e.g. Caliskan et al., 2017
→ Policy preference estimation (e.g. Laver et al., 2003)

Why text?

Text data is

- → ubiquitous
- → easily collectable
- → informative, even where other behaviour is not

Why text?

Text data is

- → ubiquitous
- → easily collectable
- → informative, even where other behaviour is not

Nevertheless also

- → awkward to work with
- → often strategically generated (or *not* generated)
- → difficult to compare across genres, languages, institutions

Not (just) NLP

Overlapping NLP tasks

- → Segmentation / tokenization: Locating words and sentences
- → Part of Speech (POS) tagging: Associating grammatical roles with words (noun, verb, determiner, preposition, etc.)
- → Parsing: grammatical structure from sentences

Distinctly NLP tasks

- → Named Entity Recognition (NER): Identifying people, places, and things
- → Information Extraction (IE): Extracting ‘facts’ (who did what to whom, when)

Difficult data

The Zipf-Mandelbrot law (Mandelbrot, 1966; Zipf, 1932)

$$C(W_i) \propto 1/\text{rank}(W_i)^{\alpha}$$

where rank(.) is the frequency *rank* of a word in the vocabulary and $\alpha \approx 1$

(This is a Pareto distribution in disguise)

Difficult data

The Zipf-Mandelbrot law (Mandelbrot, 1966; Zipf, 1932)

$$C(W_i) \propto 1/\mathrm{rank}(W_i)^{\alpha}$$

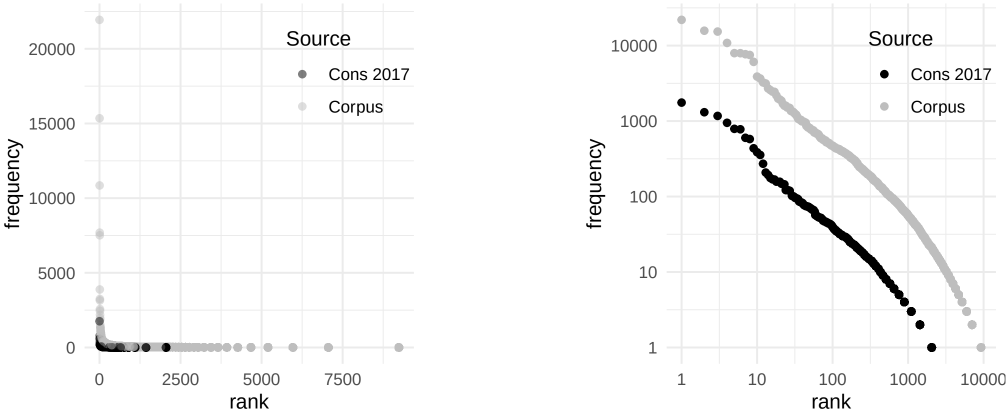
where rank(.) is the frequency *rank* of a word in the vocabulary and $\alpha \approx 1$

(This is a Pareto distribution in disguise)

Intuition:

→ Most words occur in *very* low frequencies, while a handful dominate

Difficult at all scales



This is a *power law* relationship: see also Chater and Brown (1999) on scale invariance.

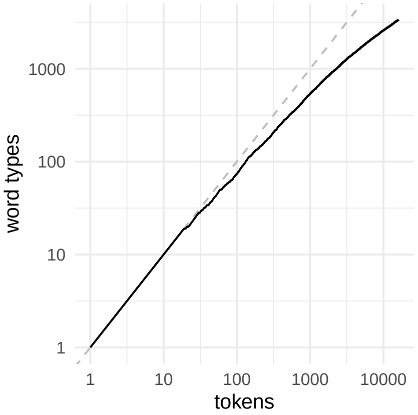
Types and Tokens

More generally: the Heaps-Herdan Law states that the number of word types appearing for the first time after n tokens is

$$D(n) = Kn^{\beta}$$

where K is between 10 and 100 and $\beta \approx 0.5$ for English.

(All the party manifestos shown here)



Frequency and interestingness

Frequency is inversely proportional to substantive interestingness

	Word	Freq.
1	the	21939
2	and	15747
3	to	15347
4	of	10850
5	we	7943
6	will	7930

Top 10

	Word	Freq.
16078	1.83	1
16079	2.20	1
16080	1.35	1
16081	33.34	1
16082	1.71	1
16083	rigination	1

Bottom ten

	Word	Freq.
20	people	1929
26	new	1507
27	government	1493
33	support	1212
34	work	1143
36	uk	1058

Top ten minus *stopwords*

Stopwords

Stopwords are a list of words that we think won't be worth keeping track of and will only get in the way of analysis

- → Not outliers (they're usually the most common!)
- → Like speech tics and pauses in a speech transcript: 'not worth transcribing'

Removing *stopwords*, while standard in computer science, is not necessarily better…

Example:

- → Standard collections contain, 'him', 'his', 'her' and 'she'.
- → Words you'd want to *keep* when analyzing an abortion debates.

Reminder: 'Preprocessing' steps like this are model fitting in disguise

Bags of words

One big distinguishing feature of text as data approaches from NLP is the willingness to make *bag of words* assumptions

Formally, the BOW assumption says: words occurrences are *exchangeable*, approximately:

→ Document *content* does not depend on the order of the words

So (de Finetti, 2008) we can model words as independently generated, *conditional on* a message θ

$$P(\text{“unemployment is socially corrosive”}) = P(\{\text{corrosive, unemployment, socially, is}\})$$

$$= \int \prod_{w}^{\substack{\{\text{corrosive, is} \\ \text{unemployment} \\ \text{socially}\}}} P(W = w \mid \theta) P(\theta) \mathrm{d}\theta$$

Clearly this is a better assumption in some genres than others…

The data and the message

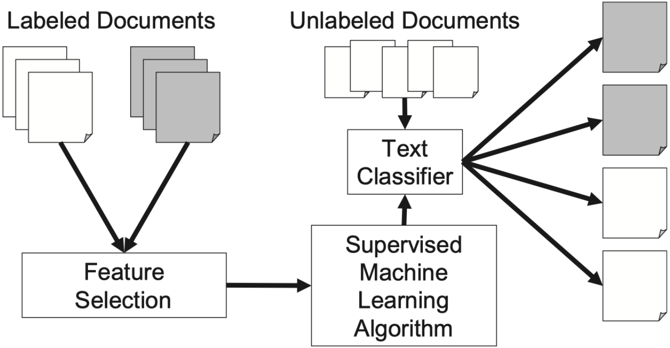
Bags of words are *contingency tables* C, or term-document / document-term / document-feature matrices, in the lingo

	corrosive	is	unemployment	socially	a		
doc 1	1	1	2	0	2	…	θ_{doc1}
doc 2	0	0	1	1	12	…	θ_{doc2}
	$\beta_{\text{corrosive}}$	β_{is}	$\beta_{\text{unemployment}}$	β_{socially}	β_{a}		

What is θ?

- → A sample of words from a *single* topic (category, subject, etc.): document classification
- → A mixed bag of *topics* (categories, emphases, etc.) in particular proportions: topic models and dictionary-based content analysis
- → A sample of words from a single *position* in some space: scaling models

Document classification



From Evans et al., 2007.

Document classification

Evan et al. try to distinguish Amicus briefs in favour of the defendants or the plaintiffs in two US affirmative action cases

- Classifier: 'Naive Bayes'

This is a *generative* classifier, meaning it tries to learn how words would be generated if you were supporting the defendant vs the plaintiff

- $P(\{W\} \mid Y = \text{plaintiffs}) = \prod_w^{\{W\}} P(W = w \mid Y = \text{plaintiffs})$
- $P(\{W\} \mid Y = \text{defendants}) = \prod_w^{\{W\}} P(W = w \mid Y = \text{plaintiffs})$

then *reverses* this using Bayes Theorem to infer

- Supports the plaintiff: $P(Y = \text{plaintiffs} \mid \{W\})$
- Supports the defendants: $1 - P(Y = \text{plaintiffs} \mid \{W\})$

Document classification

This inefficient:

- → some words are used at equal rates by both sides, so are useless for distinguishing them
- → but they're in the mix anyway, even if just noise

But conveniently, we get a vocabulary analysis as a side product, e.g.

$$\frac{P(\text{`benign'} \mid Y = \text{plaintiffs})}{P(\text{`benign'} \mid Y = \text{defendants})}$$

Intuition: If this is large then using 'benign' *distinguishes* the plaintiffs

Document classification

Term[a]	Avg. Freq. per Lib. Brief	Avg. Freq per Cons. Brief	Chi²	Interpretive Code Examples[b]
Conservative Words				
PREFER*	2.83	41.79	39.18	Proceduralist; Race/Gender Neutral Justice
BENIGN	0.07	1.17	36.14	Intent vs. Consequences; Constraint
DISCRIM*	14.86	25.04	24.13	Proceduralist; Race/Gender Neutral Justice
PURPORT*	0.44	1.88	24.13	Skepticism
CLASSIF*	2.1	11.54	22.39	Proceduralist; Race/Gender Neutral Justice
NARROW-TAILORING	0.05	0.96	19.73	Proceduralist; Strict Scrutiny
REJECT*	2.75	7.79	19.15	Oppositional Posture
JUSTIF*	2.39	12.79	18.91	Proceduralist; Constraint
FORBID*	0.38	1.63	18.91	Proceduralist; Constraint; Race/Gender Neutral Justice
PROHIBITS	0.13	0.71	18.08	Proceduralist; Constraint
RATIONALE	0.66	5.92	17.58	Proceduralist; Legalistic
AMORPHOUS	0.25	1.29	14.62	Proceduralist; Skepticism
RACE-BASED	1.08	10.46	10.59	Proceduralist; Pejorative counterpart to liberal RACE-CONSCIOUS
Liberal Words				
LEADERS	2.70	0.13	31.03	Impact; Development
WORLD	3.00	0.42	18.74	Impact; Global
NATION*	21.0	7.04	17.90	Impact; Communitarian
IMPACT*	4.13	1.04	17.49	Impact
EFFECTIVE	2.78	0.75	16.54	Impact; Effectiveness
SOCIAL	6.84	1.71	16.05	Impact; Communitarian
COMMUNIT*	8.75	1.75	15.35	Impact; Communitarian
BUSINESS*	4.56	0.58	10.28	Impact; Efficiency; Distributive Justice
DESEGREGATION	2.34	0.17	10.24	Remedial Justice
GROW*	2.38	0.33	10.24	Change; Development
WORKFORCE	1.64	0.00	9.81	Impact; Distributive Justice; Development
RACE-CONSCIOUS	7.14	1.50	7.80	Proceduralist; Euphemistic counterpart to conservative RACE-BASED

In the dictionary, ‘benign’ has a broadly positive valence. In this situation it is quite loaded in favour of the plaintiffs.

Vocabulary contrasts

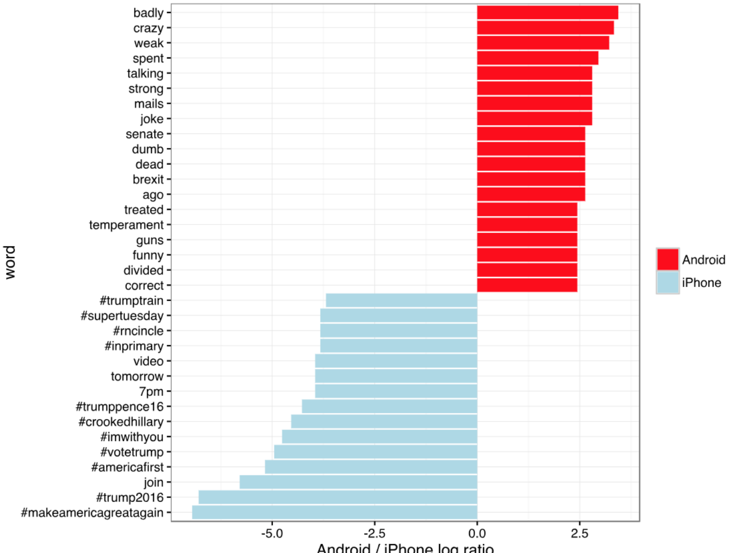
Alternatively, these comparisons may be the focus of a text analysis, not a byproduct

→ The `quanteda` package calls these differences 'keyness'

For example, here's some comparisons of the words that discriminate 2016 Trump on Twitter, depending on whether the source of the tweet is an iPhone or an Android phone

→ Theory: staff used iPhones and posted campaign messages to his account, he has an Android

ANALYZE TWITTER DATA. VOTE!



Dictionaries and topic models

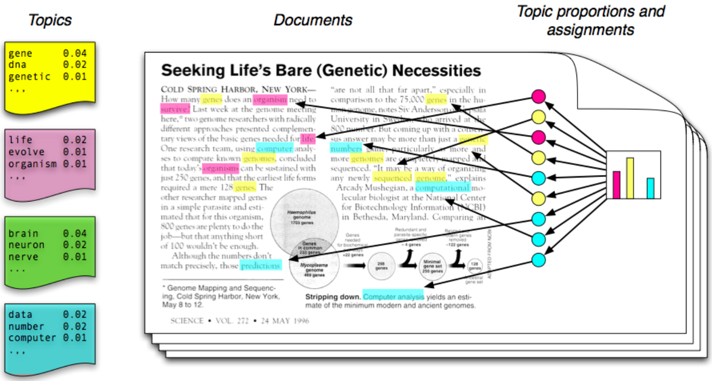
Previously we assumed that documents expressed only one thing, eg. support for the plaintiffs

What if we believed that the message was in the mixture of topics it contained?

Two approaches:

→ Confirmatory, and manual: build a content analysis dictionary
→ Exploratory (mostly), and automated: fit a topic model

Topics



From Blei et al. (2003)

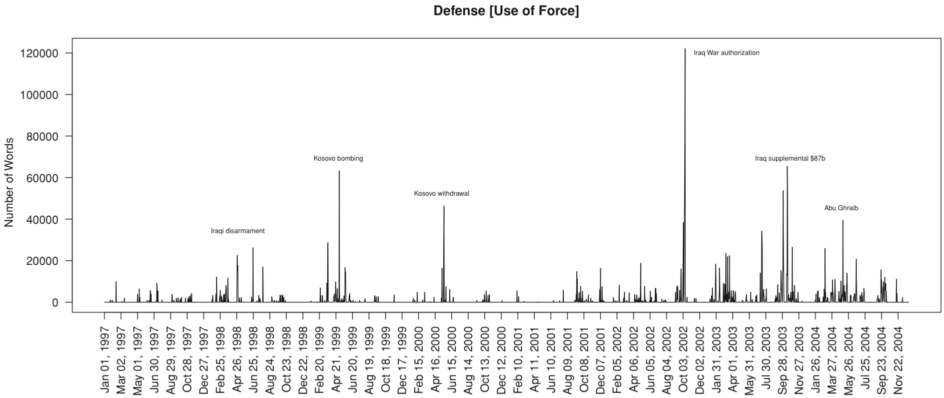
Topics

Topic (Short Label)	Keys
1. Judicial Nominations	*nomine, confirm, nomin, circuit, hear, court, judg, judici, case, vacanc*
2. Constitutional	*case, court, attornei, supreme, justic, nomin, judg, m, decis, constitut*
3. Campaign Finance	*campaign, candid, elect, monei, contribut, polit, soft, ad, parti, limit*
4. Abortion	*procedur, abort, babi, thi, life, doctor, human, ban, decis, or*
5. Crime 1 [Violent]	*enforc, act, crime, gun, law, victim, violenc, abus, prevent, juvenil*
6. Child Protection	*gun, tobacco, smoke, kid, show, firearm, crime, kill, law, school*
7. Health 1 [Medical]	*diseas, cancer, research, health, prevent, patient, treatment, devic, food*
8. Social Welfare	*care, health, act, home, hospit, support, children, educ, student, nurs*
9. Education	*school, teacher, educ, student, children, test, local, learn, district, class*
10. Military 1 [Manpower]	*veteran, va, forc, militari, care, reserv, serv, men, guard, member*
11. Military 2 [Infrastructure]	*appropri, defens, forc, report, request, confer, guard, depart, fund, project*
12. Intelligence	*intellig, homeland, commiss, depart, agenc, director, secur, base, defens*
13. Crime 2 [Federal]	*act, inform, enforc, record, law, court, section, crimin, internet, investig*
14. Environment 1 [Public Lands]	*land, water, park, act, river, natur, wildlif, area, conserv, forest*
15. Commercial Infrastructure	*small, busi, act, highwai, transport, internet, loan, credit, local , capit*
16. Banking / Finance	*bankruptci, bank, credit, case, ir, compani, file, card, financi, lawyer*
17. Labor 1 [Workers]	*worker, social, retir, benefit, plan, act, employ, pension, small, employe*

From Quinn et al. (2006)

Note: only the top most probable words are shown and topic labels are manually assigned.

Topics



From Quinn et al. (2006)

Topic model training

Topic models can be quite hard and time consuming to estimate. We start with

- A term document matrix C (the contingency table)
- A belief about the number of topics K

and try to learn

- a topic label $Z = 3$ for each word
- a 'dictionary' $\beta_{w3} = P(W = w \mid Z = 3)$ for every w and every topic
- the proportion of each topic, e.g. θ_3 in each document

These are all coupled, and all unknown.

We can help a bit with hyperparameters that give the model a 'prior' over θ and/or over β

Interpreting topic models

Ideally we'd like to be able to say: "make this one about defense"

Unfortunately, that level of high level control is an unsolved problem

We can only – after the fact – label the topics, and hope some are topics we want.

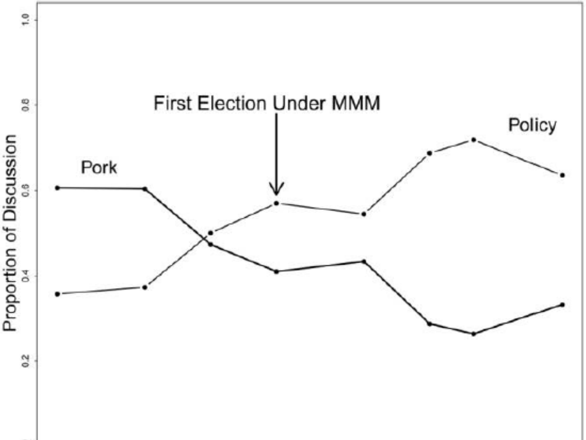
Topic model topics

Are they good, these topics? Ironically

- → the better the *statistical* properties of the model the less interpretable it tends to be (Chang et al., 2009)
- → Clearly we're missing something with the model structure…

Explaining topic prevalence

Often we want to both measure and explain the prevalence of topic mentions, e.g. the effects of a Japanese electoral reform (Catalinac, 2018)



Structural topic model

Topic models usually end by presenting us with $\hat{\theta}$ for each document and a dictionary of βs

If we like some of the topics, we might want to know how they vary with external information, e.g.

→ How does rate of topic 3, say 'defence', change with the party of the speaker?

This is a regression model with

→ Speaker party indicator X (observed)
→ proportion of the speech assigned to topic 3 as Y^* (inferred, not observed)
→ Covariates Z, e.g. committee membership, date, etc. (observed)

The *structural topic model* (Roberts et al., 2014) mixes together

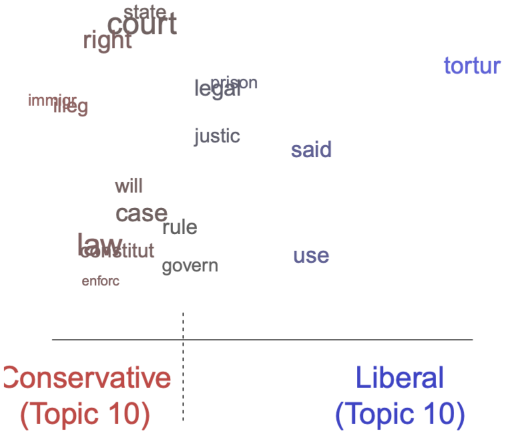
→ fitting the topic model
→ conditioning its output on the X and Z

Convenient!

Structural topic model

Having a topic model allows us to get contrast vocabulary *within topic* too.

Here's contrasting usage when talking about Guatanamo Bay in a Bush era data set



Scaling

We can also think about document living in some kind of *space* with θ as the positione.g.

- → affect, a.k.a. *sentiment analysis*
- → unidimensional policy preferences
- → multidimensional ideological position

How to place documents in space?

- → Think of a row in the document term matrix as a vocabulary profile, e.g. by normalize the counts
- → This is a point in a (very high-dimensional) space
- → Which has distances to every other document in that space

We can, and do, cluster documents this way.

Scaling

But we can also collapse them down into a smaller space, e.g. one or two dimensions

- → Often we think they really live there
- → Sometimes it's just visualization

The model is quite simple. If C_{ij} is the number of times the j-th word occurs in the i-th document, then, in one dimension

$$\log C_{ij} = \alpha_i + \psi_j + \theta_i \beta_j$$

Scaling

But we can also collapse them down into a smaller space, e.g. one or two dimensions

- → Often we think they really live there
- → Sometimes it's just visualization

The model is quite simple. If C_{ij} is the number of times the j-th word occurs in the i-th document, then, in one dimension

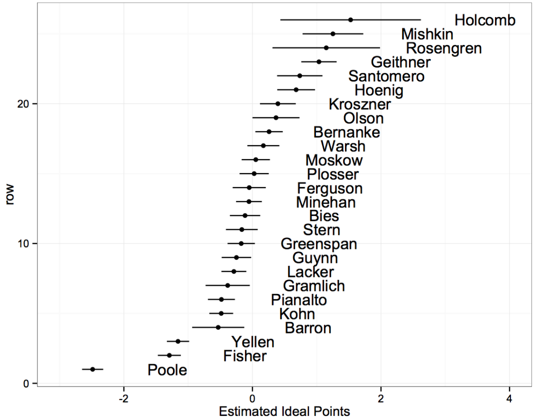
$$\log C_{ij} = \alpha_i + \psi_j + \theta_i \beta_j$$

and in more than K (orthogonal) dimensions

$$\log C_{ij} = \alpha_i + \psi_j + \sum_k^K \theta_i^{(k)} \sigma^{(k)} \beta_j^{(k)}$$

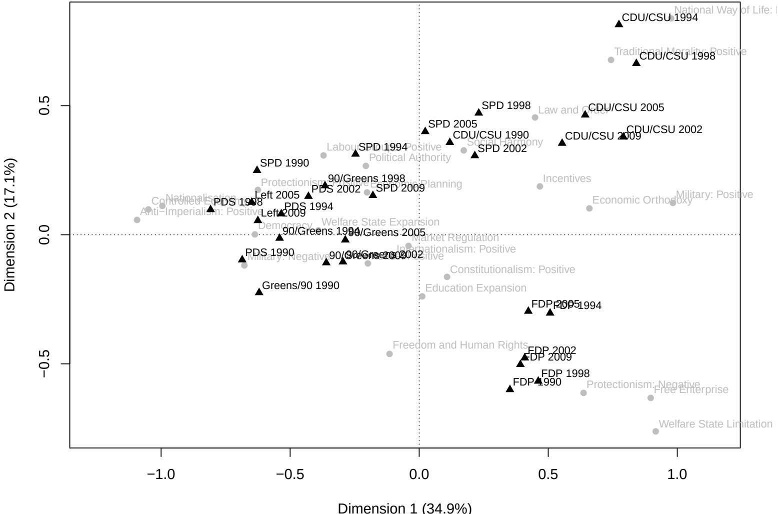
where $\sigma^{(k)}$ is the importance of that dimension to C

Scaling one dimension

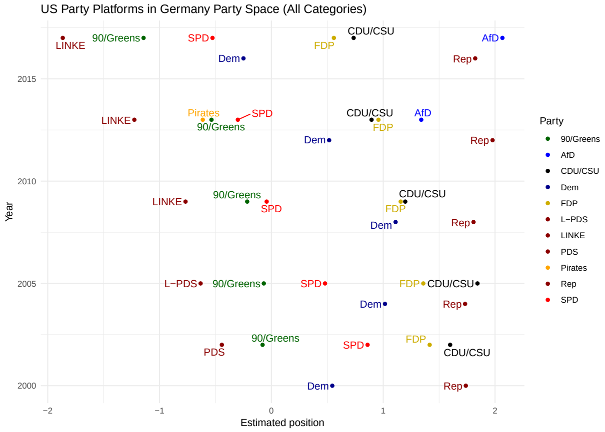


Estimated FOMC member ideal points from meeting transcripts Baerg and Lowe (2020)

Scaling several



In each other's space



Link: the pretty version at the New York Times

Text (as data)

Lots of possibilities – ask about them in class!

References

Baerg, N. & Lowe, W. (2020). 'A textual taylor rule: Estimating central bank preferences combining topic and scaling methods'. *Political Science Research and Methods*, *8*(1), 106–122.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). 'Latent dirichlet allocation'. *Journal of Machine Learning Research*, *3*, 993–1022.

Caliskan, A., Bryson, J. J. & Narayanan, A. (2017). 'Semantics derived automatically from language corpora contain human-like biases'. *Science*, *356*(6334), 183–186.

Catalinac, A. (2018). 'Positioning under alternative electoral systems: Evidence from japanese candidate election manifestos'. *American Political Science Review*, *112*(1), 31–48.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C. & Blei, D. M. (2009). 'Reading tea leaves: How humans interpret topic models'. *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 288–296.

References

Chater, N. & Brown, G. D. A. (1999). 'Scale-invariance as a unifying psychological principle.' *Cognition*, *69*(3), 817–24.

de Finetti, B. (2008). 'Philosophical lectures on probability' (A. Mura, Ed.; H. Hosni, Trans.). Springer.

Evans, M., McIntosh, W., Lin, J. & Cates, C. (2007). 'Recounting the courts? applying automated content analysis to enhance empirical legal research'. *Journal of Empirical Legal Studies*, *4*(4), 1007–1039.

Gamson, W. A. & Modigliani, A. (1989). 'Media discourse and public opinion on nuclear power: A constructionist approach'. *American Journal of Sociology*, *95*(1), 1–37.

Grimmer, J., Shorey, R., Wallach, H. M. & Zlotnick, F. (2011). 'A class of bayesian semiparametric cluster-topic models for political texts'. 1–43.

Laver, M., Benoit, K. R. & Garry, J. (2003). 'Extracting policy positions from political texts using words as data'. *American Political Science Review*, *97*(2), 311–331.

References

Mandelbrot, B. (1966). Information theory and psycholinguistics: A theory of word frequencies. In P. Lazarsfeld & N. Henry (Eds.), *Readings in mathematical social science*. MIT Press.

Mosteller, F. & Wallace, D. L. (1963). 'Inference in an authorship problem'. *Journal of the American Statistical Society*, *58*, 275–309.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H. & Radev, D. R. (2006). 'An automated method of topic-coding legislative speech over time with application to the 105th-108th us senate'. *Midwest Political Science Association Meeting*.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. & Rand, D. G. (2014). 'Structural topic models for open-ended survey responses'. *American Journal of Political Science*, *58*(4), 1064–1082.

Zipf, G. K. (1932). 'Selected studies of the principle of relative frequency in language'. Oxford University Press.