Text as Data: Summer School

William Lowe
Hertie School

July 13, 2021

TEXT AS...

From Padua (2015)

Plan

- → Text as data
- → Document classification
- → Models for topicful documents
- → Documents in space

TEXT AS DATA

Your instructor

- → Dr William Lowe
  conjugateprior@gmail.com
- → Senior Research Scientist
  Data Science Lab, Hertie School
- → The emergency backup instructor
  "in case of emergency break class"

Behind the curtain

- → Dr Olga Gasparyan
- → Huy Ngoc Dang
- → Bruno Ponne

TEXT AS DATA

Your instructor

- → Dr William Lowe
  conjugateprior@gmail.com
- → Senior Research Scientist
  Data Science Lab, Hertie School
- → The emergency backup instructor
  "in case of emergency break class"

Behind the curtain

- → Dr Olga Gasparyan
- → Huy Ngoc Dang
- → Bruno Ponne

Materials

- → Practical exercises are available as zip file on the course page
- → Each session has a folder
- → Each folder contains an RStudio project file (click to launch)
- → `*.html` is a code walk-through
- → `*.R` is the code
- → You'll never need to change the working directory

TEXT AS DATA

Broad approaches to studying text data

- → Just read it and think a bit, e.g. op-eds, punditry, kremlinology, grand strategy, etc.
- → Discourse Analysis
- → Natural Language Processing (NLP)
- → Text as Data (TADA)

the last two are, broadly, Computational Linguistics, but with a different focus

NOT DISCOURSE ANALYSIS

> *Although discourse analysis can be applied to all areas of research, it cannot be used with all kinds of theoretical framework. Crucially, it is not to be used as a method of analysis detached from its theoretical and methodological foundations. Each approach to discourse analysis that we present is not just a method for data analysis, but a theoretical and methodological whole - a complete package. [...] In discourse and analysis theory and method are intertwined and researchers must accept the basic philosophical premises in order to use discourse analysis as their method of empirical study.*
>
> *(Jørgensen & Phillips, 2002)*

Apparent differences are theoretical. The important difference for us is that

→ Discourse analysis *tightly couples* theory and measurement

Substantive theory ≠ textual measurement…but they do have implications for one another

NOT (JUST) NLP

A typical NLP pipeline

- → *Segmentation / tokenization*
- → Part of Speech (POS) tagging
- → Parsing
- → Named Entity Recognition (NER)
- → Information Extraction (IE)

NOT (JUST) NLP

A typical NLP pipeline

- → *Segmentation / tokenization*
- → Part of Speech (POS) tagging
- → Parsing
- → Named Entity Recognition (NER)
- → Information Extraction (IE)

中国国家主席习近平在对美国的首次国事访问中，周二晚上展示了他对美国历史和流行文化的熟悉。

President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

NOT (JUST) NLP

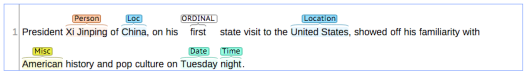
A typical NLP pipeline

- → Segmentation / tokenization
- → *Part of Speech (POS) tagging*
- → *Parsing*
- → *Named Entity Recognition (NER)*
- → Information Extraction (IE)

Tools:

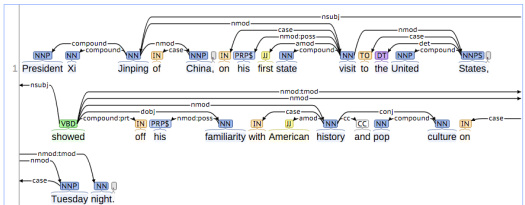
- → Spacy (spacy.io) and accessible from R using {spacyr}
- → Stanford NLP tools (nlp.stanford.edu)

Named Entity Recognition:



Coreference:

Basic Dependencies:



TEXT AS DATA: THE APPROACH

We are the measurement component for social science theory

- → Theory provides the things to be measured
- → Words and sometimes other things provide the data to measure them
- → Language agnostic, evidentially behaviourist, structurally indifferent, shamelessly opportunistic
- → obsessed with counting words

If Discourse Analysis offers close reading, we will offer *distant reading*

Advantages

- → Scales well
- → Easy to integrate into existing models
- → Can guide close reading later

TRANSCENDENTAL QUESTION

What are the *conditions for the possibility* for taking a TADA approach

In plainer language:

→ How could this possibly work?

An uncharacteristically dashing Kant

BIG PICTURE

There is a *message* or *content* that cannot be directly observed, e.g.

- → the topic of this talk
- → my position on some political issue
- → the importance of defence issues to a some political party

and *behaviour*, including *linguistic behaviour*, e.g.

- → yelling, writing, lecturing

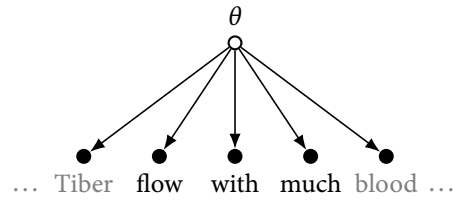
which *can* be directly observed.

Although language can do things directly – inform, persuade, demand, threaten, (Austin, 1962) – we'll focus on its signal properties: *expressed message* and its *words*…

COMMUNICATION

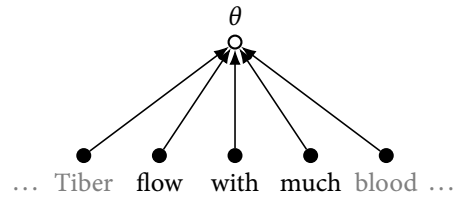
To *communicate* a message θ (or Z) to a producer (the speaker or writer) *generates* words of different kinds in different quantities

For models: the *generative* mode



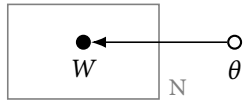
To *understand* a message the consumer (the hearer, reader, coder) uses those words to *reconstruct* the message

For models: the *discriminative* mode



NOTATION: LOOKING AHEAD

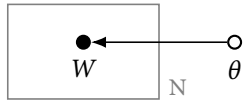
We'll represent the sample of N words in a 'plate'



And usually use Z as a categorical 'message' and θ as a continuous one

NOTATION: LOOKING AHEAD

We'll represent the sample of N words in a 'plate'



And usually use Z as a categorical 'message' and θ as a continuous one

We can read this several ways:

- → causal: θ causes those words to be generated
- → statistical (general): There exists a conditional distribution of W given θ
- → statistical (measurement): Ws are *conditionally independent* given θ
- → practical: Somewhere in the model is a table relating θ to W

COMMUNICATION

This process is

- → stable (Grice, 1993; Searle, 1995)
- → conventional (Lewis, 1969/1986)
- → disruptible (Riker et al., 1996)
- → empirically underdetermined (Davidson, 1985; Quine, 1960)

How to model this without having to solve the problems of linguistics (psychology, politics) first?

Rely on:

- → instrumentality
- → reflexivity
- → randomness

CONVENTION

TOP DEFINITION

beverage

What **drinks** are when they're not **allowed** **somewhere**.

Sign: *"No food or* ***beverages****"*

(Urban dictionary, 12 July 2021)

The difference between

- → X means Y
- → X is used to mean Y

COMMUNICATION: REFLEXIVITY

Politicians are often nice enough to talk as if they really do communicate this way

> *My theme here has, as it were, four heads. [...] The first is articulated by the word "opportunity" [...] the second is expressed by the word "choice" [...] the third theme is summed up by the word "strength" [and] my fourth theme is expressed well by the word "renewal".*

(Note however, these words occur 2, 7, 2, and 8 times in 4431 words)

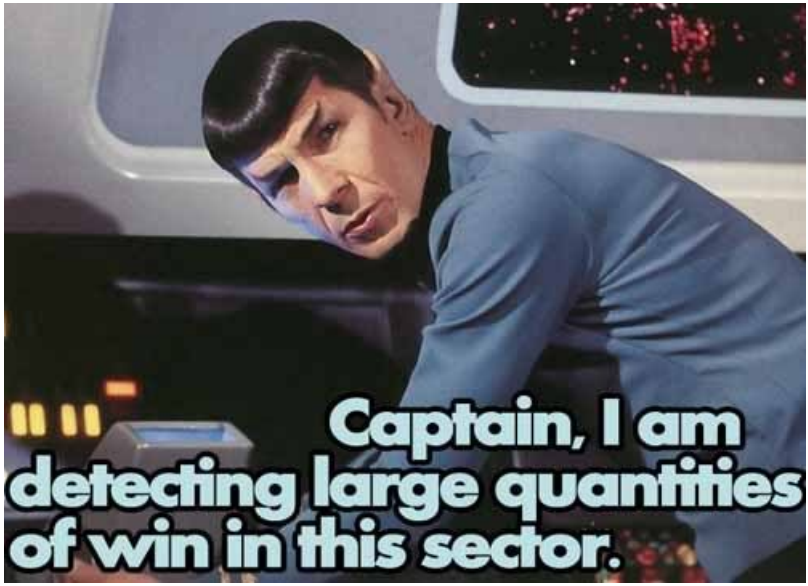
COMMUNICATION: REFLEXIVITY

Politicians are often nice enough to talk as if they really do communicate this way

> *My theme here has, as it were, four heads. [...] The first is articulated by the word "opportunity" [...] the second is expressed by the word "choice" [...] the third theme is summed up by the word "strength" [and] my fourth theme is expressed well by the word "renewal".*

(Note however, these words occur 2, 7, 2, and 8 times in 4431 words)

> *A couple months ago we weren't expected to win this one, you know that, right? We weren't...Of course if you listen to the pundits, we weren't expected to win too much. And now we're winning, winning, winning the country – and soon the country is going to start winning, winning, winning.*


Captain, I am
detecting large quantities
of win in this sector.

COMPARISON PROBLEMS

Quantitative text analysis works best when language usage is stable, conventionalized, and instrumental.

Implicitly, that means *institutional language*, e.g.

- → courts
- → legislatures
- → op-eds
- → financial reporting

Institution-specificity analyses inevitably create *comparability* problems, e.g. between

- → upper vs lower chamber vs parliamentary hearings
- → bureaucracy vs lobby groups (Klüver, 2009)
- → European languages (Proksch et al., 2019)

RHETORICAL INSTABILITY

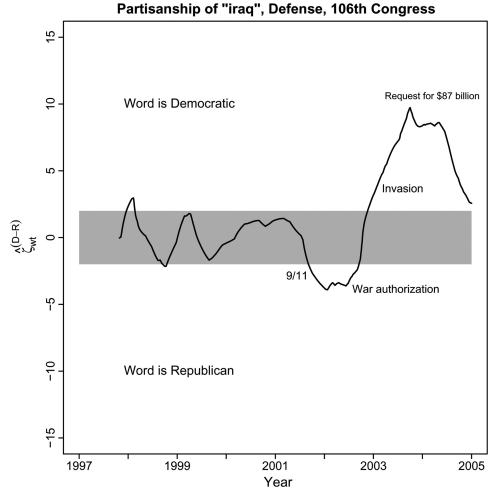
We are going to design instruments to measure θ and are going to assume that the $\theta \longrightarrow W$ relationships are institutionally stable

What if they aren't?

RHETORICAL INSTABILITY

We are going to design instruments to measure θ and are going to assume that the $\theta \longrightarrow W$ relationships are institutionally stable

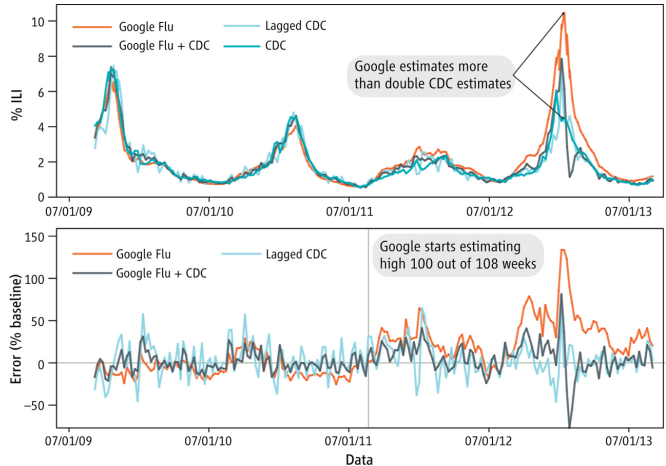
What if they aren't?



MEASUREMENT INSTABILITY

Google flu trends. Predictive, until it wasn't.

(Lazer et al., 2014)



REFLEXIVE SOLUTIONS

Sometimes actors are happy to solve comparability problems for us, e.g.

- → Lower court opinions (Corley et al., 2011) or Amicus briefs (Collins et al., 2015) *embedded in* Supreme Court opinions
- → ALEC model bills *embedded in* state bills (Garrett & Jansa, 2015)

A perfect jobs for *text-reuse* algorithms…

COMMUNICATION: RANDOMNESS

Why randomness?

- → You almost never *say exactly the same words twice*, even when you haven't changed your mind about the message.
- → so words are the result of some kind of *sampling process*.
- → We model this process as random because we *don't know or care* about all the causes of variation

COMMUNICATION: RANDOMNESS

Why randomness?

- → You almost never *say exactly the same words twice*, even when you haven't changed your mind about the message.
- → so words are the result of some kind of *sampling process*.
- → We model this process as random because we *don't know or care* about all the causes of variation

Note:

- → What is 'signal' and what is 'noise' is relative to your and the sources' purposes

COMMUNICATION: RANDOMNESS

Why randomness?

- → You almost never *say exactly the same words twice*, even when you haven't changed your mind about the message.
- → so words are the result of some kind of *sampling process*.
- → We model this process as random because we *don't know or care* about all the causes of variation

Note:

- → What is 'signal' and what is 'noise' is relative to your and the sources' purposes

Also, we're all secretly Bayesians

WORDS AS DATA

What do we know about words as data?

They are *difficult*

→ High dimensional

→ Sparsely distributed (with skew)

→ Not equally informative

DIFFICULT WORDS

Example: Conservative party 2017 manifesto compared to other parties over four elections:

- → *High dimensional*. 3784 word types (adult native english speakers know 20-35,000)
- → *Sparse*. That's about 23.5% the 16,083 word types deployed over these elections
- → *Skewed*. Of these 1731 words appeared *exactly once* and the most frequent word 1757 times

DIFFICULT WORDS

Example: Conservative party 2017 manifesto compared to other parties over four elections:

- → *High dimensional*. 3784 word types (adult native english speakers know 20-35,000)
- → *Sparse*. That's about 23.5% the 16,083 word types deployed over these elections
- → *Skewed*. Of these 1731 words appeared *exactly once* and the most frequent word 1757 times

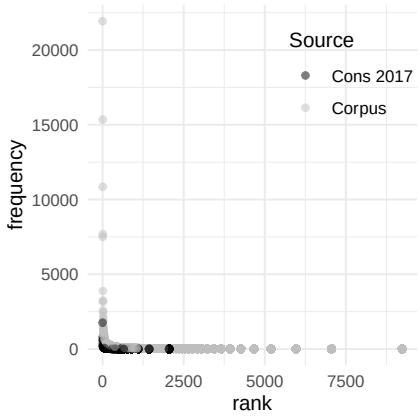
More generally: the Zipf-Mandelbrot law (Mandelbrot, 1966; Zipf, 1932)

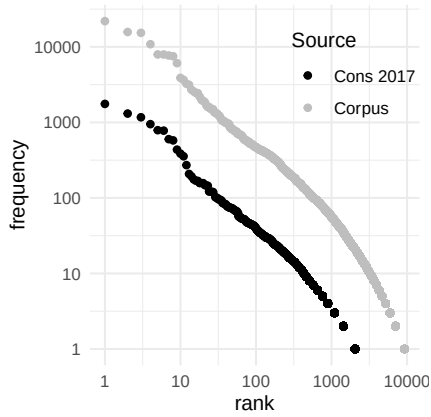
$$F(W_i) \propto 1/\text{rank}(W_i)^{\alpha}$$

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

This is a Pareto distribution in disguise

DIFFICULT AT ALL SCALES





See Chater and Brown (1999) on scale invariance.

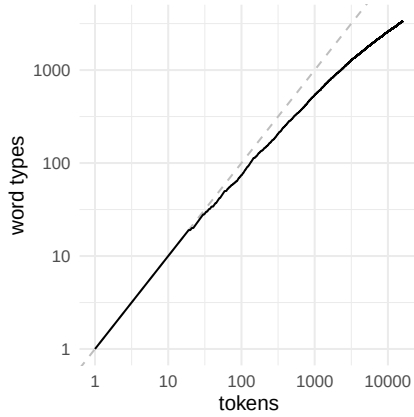
TYPES AND TOKENS

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*

DEALING WITH DIFFICULT WORDS

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.

DEALING WITH DIFFICULT WORDS

For large amounts of text summaries are not enough.

We need a *model* to provide assumptions about

- → *equivalence*
- → *exchangeability*

Text as data approaches started off asserting equivalences, and ended up modeling with increasingly sophisticated versions of exchangeability

Since ontogeny recapitulates phylogeny, let's walk through some standard text processing steps, asserting equivalences along the way…

PUNCTUATION INVARIANCE

As I look ahead I am filled with foreboding. Like the Roman I seem to see 'the river Tiber flowing with much blood'…"

(Powell, 1968)

PUNCTUATION INVARIANCE

As I look ahead I am filled with foreboding. Like the Roman I seem to see 'the river Tiber flowing with much blood'…"

(Powell, 1968)

index	token
1	as
2	i
3	look
4	ahead
5	i
6	am
7	…

index	token
1	like
2	the
3	roman
4	i
5	seem
6	to
7	…

LEXICAL UNIVOCALITY

type	count
as	1
i	2
look	1
ahead	1
am	1
…	…

token	count
like	1
the	1
roman	1
i	1
seem	1
to	1
…	…

ORDER INVARIANCE

		‘doc’ 1	‘doc’ 2
type	ahead	1	0
	am	1	0
	as	1	0
	i	2	1
	like	0	1
	look	1	0
	roman	0	1
	seem	0	1
	the	0	1
	to	0	1
	…	…	…

This is the notorious bag-of-words or *exchangeability* assumption

COUNT DATA

We have turned a corpus into a *contingency table*.

→ Or a term-document / document-term / document-feature matrix, in the lingo

	ahead	am	i	like	look		
doc 1	1	1	2	0	1	…	θ_{doc1}
doc 2	0	0	1	1	0	…	θ_{doc2}
	β_{ahead}	β_{am}	β_{i}	β_{like}	β_{look}		

Everything you learned in your last categorical data analysis course applies here

→ except that the parts of primary interest are *not observed*

STATISTICAL ASSUMPTIONS

So what are we going to assume about the word counts?

Word counts/rates are conditionally *Poisson*:

$$W_j \sim \text{Poisson}(\lambda_j)$$
$$\text{E}[W_j] = \text{Var}[W_j] = \lambda_j$$

We'll let model assumptions determine how λ is related to θ

→ typically generating proportional increases or decreases in λ

STATISTICAL ASSUMPTIONS ABOUT WORDS

The Poisson assumption implies that for conditional on document length, word counts are *Multinomial*:

$$W_{i1} \dots W_{iV} \sim \text{Mult}(W_{i1} \dots W_{iV} \mid \pi_1 \dots \pi_V, N_i)$$

Here

$$\text{E}[W] = N\pi$$

and

$$\text{Cov}[W_j, W_k] = -N_i \pi_j \pi_k$$

Negative covariance is due to the 'budget constraint' N_i

ASIDE: ABSENCE

Statistical models of text deal with (some kinds of) *absence* as well as presence

We will be concerned with two kinds of absence:

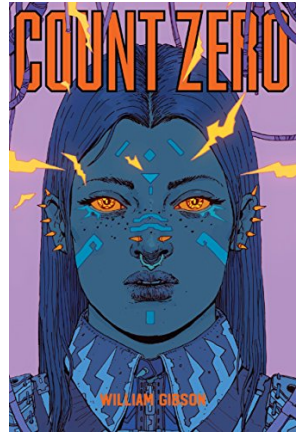
- → Not seeing a word used – a 'zero count'
- → Not seeing a document at all – 'sample selection'

(roughly overlapping with item vs unit non-response)

ASIDE: ABSENCE OF WORDS

Not seeing a *word* used is fairly easy to deal with

- → Zero counts are just counts that happen to be zero
- → Absence is informative to the extent it is surprising
- → Surprise implies expectations, and expectations imply a model



ASIDE: ABSENCE OF TEXT

Not seeing a *document* is harder.

→ What documents could we have seen but did not

→ What would we have inferred about content had we seen them?

Proksch and Slapin (2014) is a formal and treatment of this problem for legislative debate (see also Giannetti & Pedrazzani, 2016)

→ institutionally specific, because sample selection is a *research design problem*

UNITS OF ANALYSIS

Conventionally, text comes in the 'documents' and contains 'words', but these are *terms of art*.

You choose what is a document

- → documents
- → chapters
- → sections
- → window contexts
- → sentences
- → tweets
- → responses

You choose what is a word

- → contiguous letters separated by white space
- → lemmas / stems
- → bigrams and n-grams
- → phrases and names
- → mentions of topics
- → expressions of positive affect

Anything we can *count*, really…

UNITS OF ANALYSIS

General advice:

- → Let the substance guide
- → Keep your options open; whether a model is realistic is relative to purpose

Technical constraints:

- → Some unit choices will enable (or rule out) certain models
- → Some bags of words are *baggier* than others

MODEL DECISIONS

For each research problem involving text analysis we need to ask:

- → What *structure* does θ or Z have? topic, topic proportions, position
- → What is *observed*, *assumed*, and *inferred*?
- → *relationship* between θ or Z and the words?

Which direction do we want to model?

- → Discriminative
- → Generative

Discriminative

We sometimes see Z or θ and can learn

$$P(\theta \mid W_1 \dots W_N)$$

from a corpus. Typically *confirmatory*

Generative

We don't see Z or θ but can make assumptions about how words are generated from them

$$P(\theta \mid W_1 \dots W_N) = \frac{P(W_1 \dots W_N \mid \theta) P(\theta)}{P(W_1 \dots W_N)}$$

Typically *exploratory*

TEXT AS DATA

Zoom fatigue version

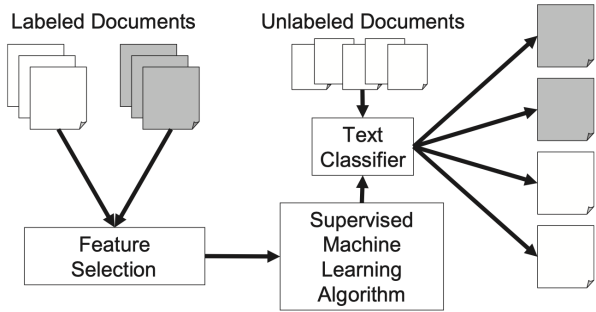
- → Text as data approaches to text analyses rely on *institutionalized* language usage,
- → They assume *stable* meaning-word relations,
- → You to decide what a document or word *is*,
- → Text's skewed high-dimensional nature is solved by with models
- → Models may be discriminative or generative

DOCUMENT CLASSIFICATION

Every document is on one of K topics / categories

We have a labeled 'training' sample

What are the rest about?



DOCUMENT CLASSIFICATION

Two sides of the one technology

- → Tool for assigning topics to new document on the basis of labeled existing documents
- → Tool for learning about how documents express topics in words

The first can be a useful research assistant

- → We want the best classifier you can train. period

The second can generate insight

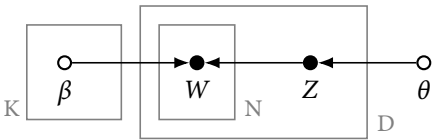
- → We want the most interpretable parameters

Sometimes we can have these together. But not often…

We'll look at *Naive Bayes*, an old but serviceable *generative* model, and its alter ego a purely *discriminative* model

NAIVE BAYES DOCUMENT CLASSIFICATION

D documents, each on topic $Z = k$ of K



This model is written *generatively*

→ How to generate words in a document on one topic

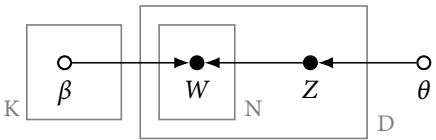
We will

→ learn these relationships

→ *update* our view of θ with new documents

NAIVE BAYES DOCUMENT CLASSIFICATION

D documents, each on topic $Z = k$ of K



This model is written *generatively*

→ How to generate words in a document on one topic

We will

→ learn these relationships

→ *update* our view of θ with new documents

GENERATION:

The proportion of documents of topic k is

$$P(Z = k) = \theta_k$$

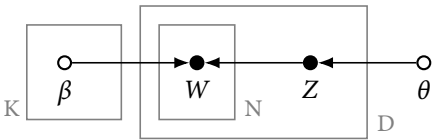
we have a *prior* over this

Then we'll *estimate* the probability that topic k generates the ith word

$$\beta_{ik} = P(W_i \mid Z = k, \beta)$$

NAIVE BAYES DOCUMENT CLASSIFICATION

D documents, each on topic $Z = k$ of K



This model is written *generatively*

→ How to generate words in a document on one topic

We will

→ learn these relationships

→ *update* our view of θ with new documents

Discriminatory mode

Of more interest is the topic of some particular document $\{W\}$

$$P(Z = k \mid \{W\}, \beta)$$

Infer this reversing the generation process with *Bayes theorem*

EXAMPLE: AFFIRMATIVE ACTION

Example application: Evans et al. (2007) attempt to

- → *Discriminate* the amicus briefs from each side of two affirmative action cases: *Regents of the University of California v. Bakke* (1978) and *Grutter/Gratz v. Bollinger* (2003).
- → *Characterize* the language used by each side

We will label the Plaintiff as 'Conservative' and the Respondents as 'Liberal'

> *All told, Bakke included 57 amicus briefs (15 for the conservative side and 42 for liberals) and Bollinger received 93 (19 conservative and 74 liberal).*
>
> *(Evans et al., 2007)*

The four briefs of Plaintiffs and Respondents formed the 'training data'

CLASSIFICATION

The document category is $Z \in \{\text{Lib}, \text{Con}\}$

$$P(Z) = \theta \qquad \textit{Prior probability}$$

$$P(\{W\} \mid Z) = \prod_j P(W_j \mid Z) \qquad \textit{The naive part}$$

Words are assumed to be generated *independently* given the category Z

$$P(\text{‘Affirmative Action’} \mid Z = \text{‘Lib’}) = P(\text{‘Affirmative’} \mid Z = \text{‘Lib’}) P(\text{‘Action’} \mid Z = \text{‘Lib’})$$

CLASSIFICATION

The document category is $Z \in \{\text{Lib}, \text{Con}\}$

$$P(Z) = \theta \qquad \textit{Prior probability}$$

$$P(\{W\} \mid Z) = \prod_j P(W_j \mid Z) \qquad \textit{The naive part}$$

Words are assumed to be generated *independently* given the category Z

$$P(\text{'Affirmative Action'} \mid Z = \text{'Lib'}) = P(\text{'Affirmative'} \mid Z = \text{'Lib'})P(\text{'Action'} \mid Z = \text{'Lib'})$$

Classification here means doing something with

$$P(Z = \text{'Lib'} \mid \{W\})$$

Strictly speaking, this is just probability estimation; classification is a separate decision problem

FITTING THE MODEL

Prior:

Estimating

$$P(Z = \text{'Lib'}) = 1 - P(Z = \text{'Con'})$$

is straightforward:

→ Count the number of 'Lib' documents and divide by the total number of documents

Likelihood

Estimating

$$P(W_j \mid Z = \text{'Lib'})$$

is also straightforward (though see McCallum & Nigam, 1993)

→ Compute the proportion of words in 'Lib' training documents that were word j

NAIVE BAYES

Posterior

Use Bayes theorem to get the probability of, e.g. an amicus brief being ‘Lib’ *given* the words inside

NAIVE BAYES

Posterior

Use Bayes theorem to get the probability of, e.g. an amicus brief being 'Lib' *given* the words inside

$$P(Z = \text{'Lib'} \mid \{W\}) = \frac{\prod_j P(W_j \mid Z = \text{'Lib'}) P(Z = \text{'Lib'})}{\prod_j P(W_j \mid Z = \text{'Lib'}) P(Z = \text{'Lib'}) + \prod_j P(W_j \mid Z = \text{'Con'}) P(Z = \text{'Con'})}$$

Oof.

It will be easier to look at *how much more likely* the brief is to be 'Lib' than 'Con'

NAIVE BAYES

$$\frac{P(Z = \text{`Lib'} \mid \{W\})}{P(Z = \text{`Con'} \mid \{W\})} = \prod_j \frac{P(W_j \mid Z = \text{`Lib'})}{P(W_j \mid Z = \text{`Con'})} \times \frac{P(Z = \text{`Lib'})}{P(Z = \text{`Con'})}$$

Every new word adds a bit of information that re-adjusts the conditional probabilities

- → Multiply by something greater than one: more ‘Lib’
- → Multiply by something less than one: more ‘Con’

UBIQUITOUS LOG RATIOS

It's often more useful to work with *logged ratios* of counts and proportions a.k.a. 'logits'

$$\log(2) \approx 0.69$$
$$\log(0.5) \approx -0.69$$

Advantages:

- → symmetrical
- → interpretable zero point
- → proportional / percentage increases and decreases
- → Psychophysical and decision-theoretic motivations (see Zhang & Maloney, 2012)
- → Measurement theoretic motivations (Rasch, IRT, Bradley Terry models etc.)
- → Makes products into additions

NAIVE BAYES IN LOGS

$$\log \frac{P(Z = \text{`Lib'} \mid \{W\})}{P(Z = \text{`Con'} \mid \{W\})} = \sum_j \log \frac{P(W_j \mid Z = \text{`Lib'})}{P(W_j \mid Z = \text{`Con'})} + \log \frac{P(Z = \text{`Lib'})}{P(Z = \text{`Con'})}$$

Every new word *adds a bit of information* that re-adjusts the conditional probabilities

TINY EXAMPLE

Example: Naive Bayes with only word class 'discriminat*'.

Assume that liberal and conservative supporting briefs are equally likely (true in the training set)

$$\frac{P(Z = \text{'Lib'})}{P(Z = \text{'Con'})} = 1$$

and

$$P(W = \text{'discriminat*'} \mid Z = \text{'Lib'}) = (26 + 13)/(20002 + 18722) \approx 0.001$$
$$P(W = \text{'discriminat*'} \mid Z = \text{'Con'}) = (70 + 48)/(17368 + 17698) \approx 0.003$$

Posterior probability ratio is about 3 to 1 in favour of the document supporting the conservative side

CONSERVATIVE VOCABULARY

Term[a]	*Avg. Freq. per Lib. Brief*	*Avg. Freq per Cons. Brief*	*Chi*[2]	*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 VOCABULARY

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

→ There are no identifiable *uniquely* partisan words

→ but these associations are stable in cases 28 years apart

REAL DISCRIMINATION

Amicus brief from ‘King County Bar Association’ containing 3667 words and 4 matches to disciminat*.

```
      that "the state shall not [discriminate] against, or grant preferential treatment
the lingering effects of racial [discrimination] against minority groups in this
 remedy the effects of societal [discrimination]. Another four Justices (Stevens
      that "the state shall not [discriminate] against, or grant preferential treatment
```

EVERY GENERATIVE MODEL

The posterior probability of a document being liberal is

$$P(Z = \text{`Lib'} \mid \{W\}) = \frac{\prod_j P(W_j \mid Z = \text{`Lib'}) P(Z = \text{`Lib'})}{\prod_j P(W_j \mid Z = \text{`Lib'}) P(Z = \text{`Lib'}) + \prod P(W_j \mid Z = \text{`Con'}) P(Z = \text{`Con'})}$$

but let's do a little rearranging

EVERY GENERATIVE MODEL

The posterior probability of a document being liberal is

$$P(Z = \text{`Lib'} \mid \{W\}) = \frac{\prod_j P(W_j \mid Z = \text{`Lib'}) P(Z = \text{`Lib'})}{\prod_j P(W_j \mid Z = \text{`Lib'}) P(Z = \text{`Lib'}) + \prod P(W_j \mid Z = \text{`Con'}) P(Z = \text{`Con'})}$$

but let's do a little rearranging

$$P(Z = \text{Lib} \mid \{W\}) = \frac{1}{1 + \exp(-\eta)}$$

$$\eta = \log \frac{P(Z = \text{`Lib'})}{P(Z = \text{`Con'})} + \sum_j \log \frac{P(W_j \mid Z = \text{`Lib'})}{P(W_j \mid Z = \text{`Con'})}$$

which might remind you of a model you've seen before…

HAS A DISCRIMINATIVE ALTER EGO

Say W is 'discriminate' and it occurs $C_{\text{discriminate}} = 12$ times in some document

then we'll then add 12 lots of

$$\beta_{\text{discriminate}} = \log \frac{P(\text{discriminate} \mid Z = \text{'Lib'})}{P(\text{discriminate} \mid Z = \text{'Con'})}$$

or

$$C_{\text{discriminate}} \times \beta_{\text{discriminate}}$$

HAS A DISCRIMINATIVE ALTER EGO

Say W is 'discriminate' and it occurs $C_{\text{discriminate}} = 12$ times in some document

then we'll then add 12 lots of

$$\beta_{\text{discriminate}} = \log \frac{P(\text{discriminate} \mid Z = \text{'Lib'})}{P(\text{discriminate} \mid Z = \text{'Con'})}$$

or

$$C_{\text{discriminate}} \times \beta_{\text{discriminate}}$$

so our final discrimination function has the form

$$P(Z = \text{'Lib'} \mid \{W\}) = \frac{1}{1 + \exp(-\eta)}$$

$$\eta = \beta_0 + C_1\beta_1 + C_2\beta_2 + \ldots + C_V\beta_V$$

This is a logistic regression on the document term matrix (Jordan, 1995) a.k.a. 'Maxent'.

DISCRIMINATIVE DOCUMENT CLASSIFICATION

Naive Bayes and Logistic Regression are in some sense the 'same model'

→ As it happens, *any* exponential family choice for $P(W_j \mid Z)$ has logistic regression as its discriminative model

DISCRIMINATIVE DOCUMENT CLASSIFICATION

Naive Bayes and Logistic Regression are in some sense the 'same model'

→ As it happens, *any* exponential family choice for $P(W_j \mid Z)$ has logistic regression as its discriminative model

For easy illumination but weaker classification performance:

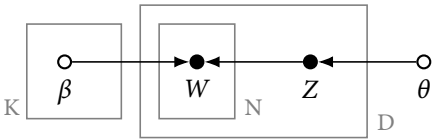
→ Naive Bayes

For less illumination but stronger classification performance:

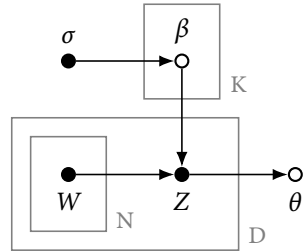
→ Regularized logit

→ or Random forests, Support Vector Machines, etc.

NAIVE BAYES AND LOGIT



Naive Bayes (generative)



Logistic regression (discriminative)

PERFORMANCE

Logistic regression is more *focused*

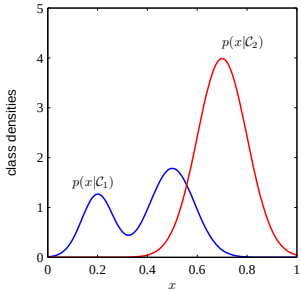
→ No interest in $P(W_1 \dots W_V)$. Words can be conditionally independent, or not. It just wants the decision boundary

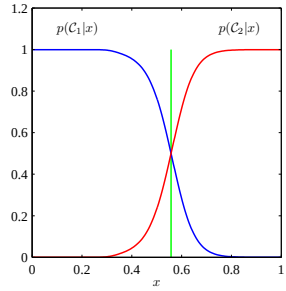
PERFORMANCE

Logistic regression is more *focused*

→ No interest in $P(W_1 \dots W_V)$. Words can be conditionally independent, or not. It just wants the decision boundary

Intuition:





NAIVE BAYES AND LOGIT

But slower and hungrier

→ β estimates converge at rate N, compared to $\log N$ for Naive Bayes' probability ratios

NAIVE BAYES AND LOGIT

But slower and hungrier

→ β estimates converge at rate N, compared to $\log N$ for Naive Bayes' probability ratios

Needs extra guidance to work well

→ We fit Naive Bayes on four documents

→ Logistic regression will requires *regularization* for that to work

Some natural regularization strategies are expressed as prior beliefs that coefficients are 'small'

→ 'Ridge regression', a.k.a. L2: $\beta_j \sim \text{Normal}(0, \sigma^2)$

→ 'Lasso', a.k.a. L1: $\sum_j |\beta_j| < \sigma$

NAIVE BAYES AND LOGIT

But slower and hungrier

→ β estimates converge at rate N, compared to $\log N$ for Naive Bayes' probability ratios

Needs extra guidance to work well

→ We fit Naive Bayes on four documents

→ Logistic regression will requires *regularization* for that to work

Some natural regularization strategies are expressed as prior beliefs that coefficients are 'small'

→ 'Ridge regression', a.k.a. L2: $\beta_j \sim \text{Normal}(0, \sigma^2)$

→ 'Lasso', a.k.a. L1: $\sum_j |\beta_j| < \sigma$

Usually better

→ Classification performance is usually better: lower bias, higher variance

→ Interpretation is trickier

THE MODEL TRADEOFF

This performance tradeoff is very general:

- → By adding bias (strong assumptions about the data) we can reduce variance
- → By adding flexibility we can reduce bias and have a more expressive model, but we'll need more and better data

The interpretation tradeoff is also general:

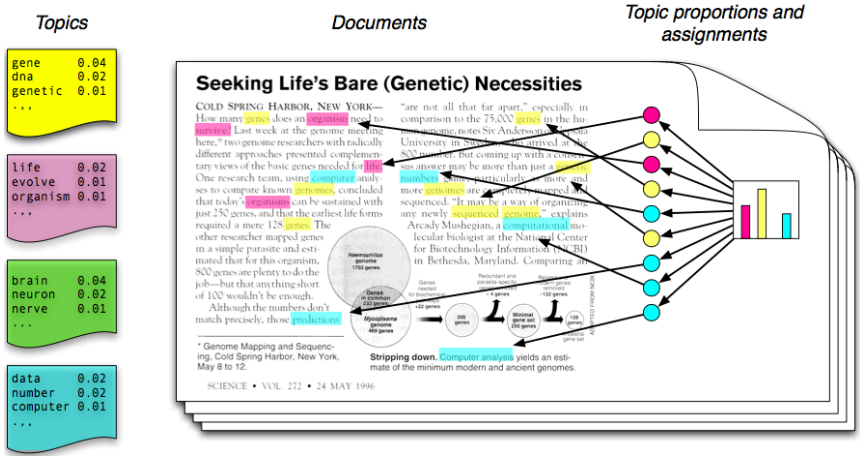
- → Better statistical performance often leads to less interpretable models (Chang et al., 2009)
- → We usually prefer the interpretable side!

TEXT AS DATA

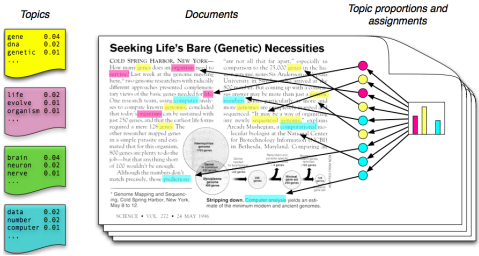
Zoom fatigue version

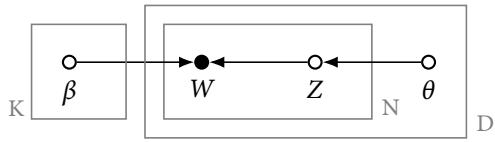
- → Document classification models assume each document has exactly one topic / category
- → Naive Bayes, a generative classifier, learns *how diagnostic* each word is for each topic
- → but may not classify so well…
- → Logistic regression (and related models, e.g. neural networks, support vector machines) is the *discriminative* version
- → but requires *regularization* to work well on text data

TOPICS IN DOCUMENTS



TOPICS IN DOCUMENTS





$$\theta = [0.062,\ 0.062,\ 0.5,\ 0.33]$$

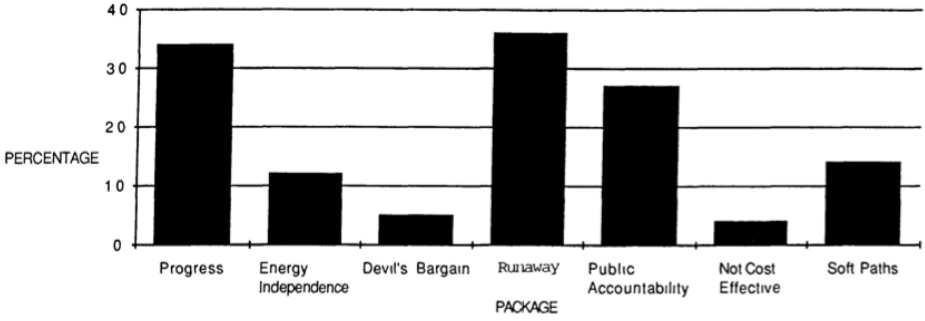
Z	2	3	4	3	3	3	4	4	4	3	3	1
W	like	the	Roman	I	see	the	River	Tiber	foaming	with	much	blood

CONNECTING TOPICAL CONTENT TO POLITICS

We're usually interested in category proportions per unit (usually document), e.g.

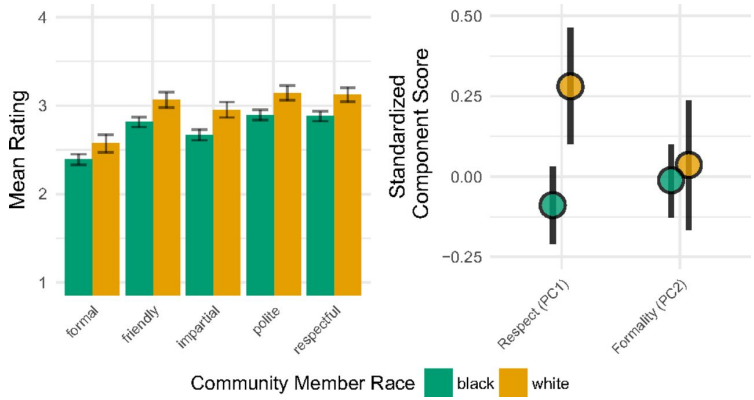
- → *How much* of this document is about national defense?
- → What is the *difference* of aggregated left and aggregated right categories (RILE)
- → How does the *balance* of human rights and national defense change over time?

TALKING LIKE A NEWSPAPER



From Gamson and Modigliani (1989)

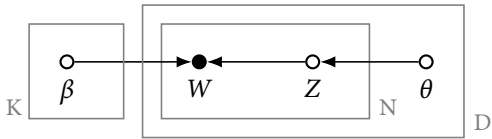
TALKING TO POLICE



From Voigt et al. (2017)

MODELS FOR TOPICFUL DOCUMENTS

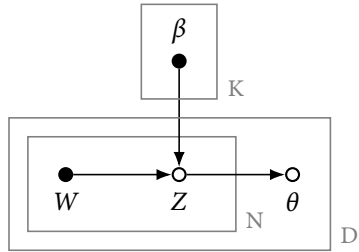
Generative version



Topic models, e.g. Latent Dirichlet Allocation

- → Learn β and θ from W
- → $\beta_{ik} = P(W_i \mid Z = k)$ for all words
- → Infer Zs

Discriminative version



Dictionary-based content analysis

- → Assert (not learn) β
- → $\beta_{ik} = P(Z = k \mid W_i, \beta) \in \{0, 1\}$
- → Infer Z and θ

DICTIONARY

Here's an excerpt from the Economy section of the dictionary in Laver and Garry (2000)

state reg	market econ
accommodation	assets
age	bid
ambulance	choice*
assist	compet*
benefit	constrain*
…	…

DICTIONARY

Here's an excerpt from the Economy section of the dictionary in Laver and Garry (2000)

state reg	**market econ**
accommodation	assets
age	bid
ambulance	choice*
assist	compet*
benefit	constrain*
…	…

$\Rightarrow$

W	P(Z = 'state reg' | W)	P(Z = 'market econ'| W)
age	1	0
benefit	1	0
…	…	…
assets	0	1
bid	0	1
…	…	…

DICTIONARY

Here's an excerpt from the Economy section of the dictionary in Laver and Garry (2000)

state reg	**market econ**
accommodation	assets
age	bid
ambulance	choice*
assist	compet*
benefit	constrain*
…	…

$\Rightarrow$

W	P(Z = 'state reg' | W)	P(Z = 'market econ'| W)
age	1	0
benefit	1	0
…	…	…
assets	0	1
bid	0	1
…	…	…

With this kind of confidence, estimating θ_k is straightforward

$$\theta_k = \frac{\sum_i^N P(Z = k \mid W_i)}{\sum_j \sum_i^N P(Z = j \mid W_i)} = \frac{\sum_i \mathbb{I}[W_i \text{ matches } k]}{\sum_i \mathbb{I}[W_i \text{ matches anything}]}$$

GENERATION

This is the $P(Z \mid W)$ is the *discrimination* (comprehension) direction

→ What does this correspond to in the *generative* direction?

GENERATION

This is the $P(Z \mid W)$ is the *discrimination* (comprehension) direction

→ What does this correspond to in the *generative* direction?

The data 'must' have been generated like this for arbitrary probabilities $a, b, c, d, \ldots$.

Robust to all kinds of generation probabilities

Because the real information is *in the zeros.*

	'state reg'	'market econ'
P(W = "age" |Z)	a	**0**
P(W = "benefit" | Z)	b	**0**
…	…	…
P(W = "assets" | Z)	**0**	c
P(W = "bid" | Z)	**0**	d
…	…	…

And *this* is where things get tricky…

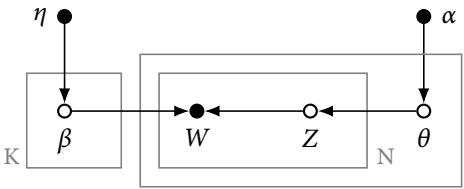
TOPIC MODELS

Turning to the generative mode…

We will try to learn θ and β,and infer Z, on the basis of W and model assumptions

→ This is a difficult problem without more constraints

We'll add them by asserting some prior expectation on the β and θ via Latent Dirichlet Allocation (Blei et al., 2003)



$$\beta_k \sim \text{Dirichlet}(\eta)$$
$$W_i \sim \text{Multinomial}(\beta_{Z_i=k}, 1)$$

$$\theta_d \sim \text{Dirichlet}(\alpha)$$
$$Z_i \sim \text{Multinomial}(\theta_d, N)$$

TOPIC MODEL TRAINING

Topic models can be quite time consuming to estimate.

→ Lots of coupled unknowns all at once

Intuition:

→ Any set of parameters make the observed word counts more or less probable

→ If we knew the Z's then estimating β and θ would be straightforward

→ If we new β and θ then estimating Z would be straightforward

→ So alternate between these steps

This simple approach is called *Gibbs sampling*

A more complete machine learning course will tell you all about it and its alternatives; we won't linger…

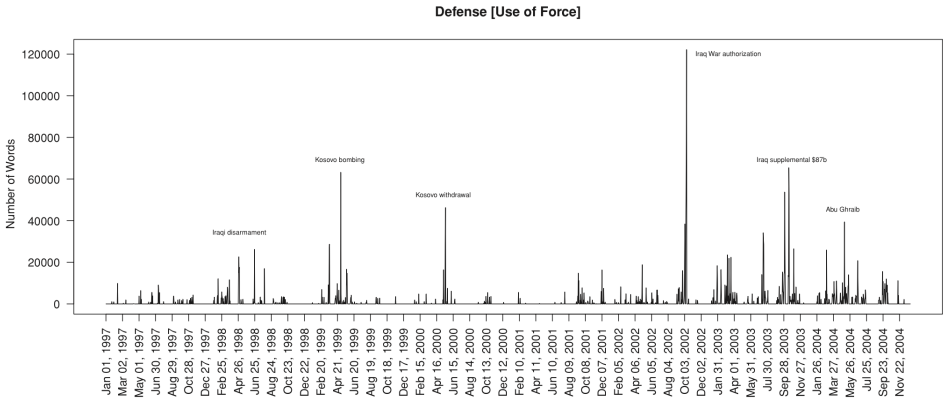
OUTPUT: β

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. (2007)

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

OUTPUT: θ_k



From Quinn et al. (2007)

INTERPRETING TOPICS

Ideally we'd like to be able to say: "make topic k about defense"

→ But we've left all the θs and βs free to vary

This level of control is an unsolved problem

→ see e.g. KeyATM, Seeded Topic Models, and a lot of other variants

We can *after the fact* assign our own labels the topics, and hope some are topics that we want.

We are fitting the exploratory form of dictionary-based content analysis

How to evaluate our new topic model?

EVALUATION

There are two main modes of evaluation:

- → Statistical
- → Human / substantive

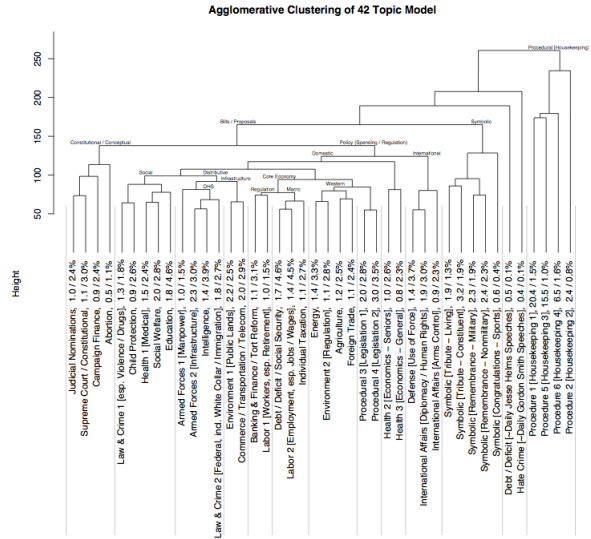
and two natural levels

- → The model as a whole: model fit, K, and topic relationships
- → Topic structure: word precision, topic coherence

Overall message: These are not yet well aligned

- → We will emphasize substance and topics

CONSTRUCT VALIDITY



Procedure:

→ Choose K

→ Fit model

→ Label topics

→ Cluster the β^k

(Quinn et al., 2007)

MODEL FIT

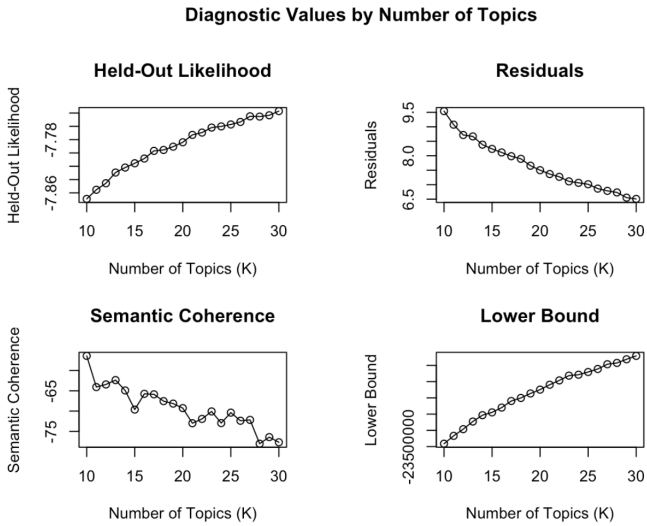
Since documents are assumed to be bags of words, then we can

- → set aside some proportion of each document
- → fit a topic model to the remainder
- → ask how probable the held out parts are under the model

The stm package calls this 'heldout likelihood by document completion'

- → Returns the average log probability of the heldout documents' words

CHOICE OF K


Diagnostic Values by Number of Topics
Held-Out Likelihood
Held-Out Likelihood
-7.86
-7.78
10
15
20
25
30
Number of Topics (K)
Residuals
Residuals
6.5
8.0
9.5
10
15
20
25
30
Number of Topics (K)
Semantic Coherence
Semantic Coherence
-75
-65
10
15
20
25
30
Number of Topics (K)
Lower Bound
Lower Bound
-23500000
10
15
20
25
30
Number of Topics (K)

CHOICE OF K

The results presented in this paper ... assume there are 43 topics present in the data. I varied the number of assumed topics from only five topics, up to 85 different topics. Assuming too few topics resulted in distinct issues being lumped together, whereas too many topics results in several clusters referring to the same issues. During my tests, 43 issues represented a decent middle ground.

(Grimmer, 2010)

CHOICE OF K

> *The results presented in this paper ... assume there are 43 topics present in the data. I varied the number of assumed topics from only five topics, up to 85 different topics. Assuming too few topics resulted in distinct issues being lumped together, whereas too many topics results in several clusters referring to the same issues. During my tests, 43 issues represented a decent middle ground.*
>
> *(Grimmer, 2010)*

We can be realists or anti-realists about topics

- → Anti-realism: topics are 'lenses'
- → Realism: topics are real discourse units, e.g. themes, categories, etc.

CHOICE OF K

> *The results presented in this paper ... assume there are 43 topics present in the data. I varied the number of assumed topics from only five topics, up to 85 different topics. Assuming too few topics resulted in distinct issues being lumped together, whereas too many topics results in several clusters referring to the same issues. During my tests, 43 issues represented a decent middle ground.*
>
> *(Grimmer, 2010)*

We can be realists or anti-realists about topics

- → Anti-realism: topics are 'lenses'
- → Realism: topics are real discourse units, e.g. themes, categories, etc.

We can *try* to be realists about the conditional independence assumption

- → Once we know the topic indicator, remaining word variation is just random → unpredictable

That's seldom true for mundane linguistic reasons

HUMANS IN THE LOOP

Chang et al. (2009) suggested two manual coded measures of precision

Precision for words

WORD INTRUSION

Choose *five* words from β^k and *one* from β^j

→ What *proportion of raters* 'agree' with the model about which word is the 'intruder'?

Proposed measure

$$\frac{1}{S}\sum_{s}^{S}\mathbb{I}[s \text{ chooses } j]$$

Topic precision

TOPIC INTRUSION

Choose

→ A snippet of text from a document

→ labels for *three* topics that have high θ for it

→ label for *one* low θ 'intruder' topic j

Raters identify i the 'intruder' topic

Proposed measure

$$\frac{1}{S}\sum_{s}^{S}\log\frac{\theta_j}{\theta_i}$$

HUMANS OUT OF THE LOOP

Precision for words

FREQUENCY

→ β^k is high

EXCLUSIVITY

→ High precision words make *well-separated* topics

$$\frac{\beta_i^k}{\sum_{j \neq k} \beta_i^j}$$

FREX

→ A weighted average of exclusivity and *frequency* (favouring exclusivity)

Precision for topics

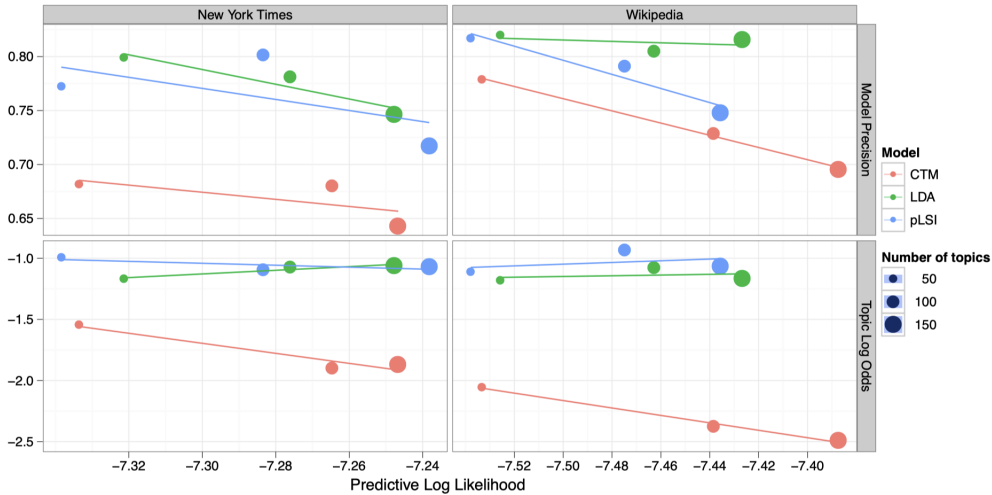
SEMANTIC COHERENCE

→ Two words that tend to appear in documents together should probably be in the same topic (Mimno et al., 2011)

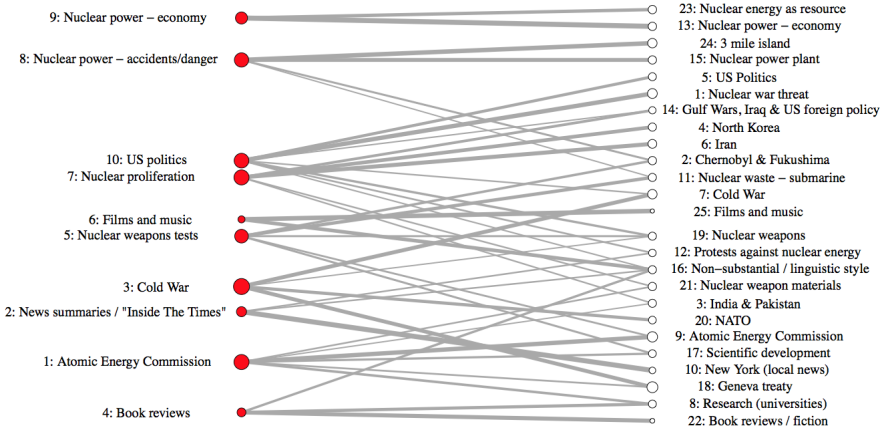
→ Computed for the M most probable words in each topic

$$\sum_i \sum_j \log \frac{D(V_i^k, V_j^k) + 1}{D(V_i^k)}$$

SEMANTIC VS STATISTICAL MEASURES



GAMSON AND MODIGLIANI REDUX



from van Atteveldt et al. (MS)

EXPLAINING TOPIC PREVALENCE

Often we want to both *measure* but also *explain*
the prevalence of topic mentions

EXPLAINING TOPIC PREVALENCE

Often we want to both *measure* but also *explain* the prevalence of topic mentions

Example: What are the effects of a Japanese house electoral reform on candidate platforms? (Catalinac, 2018)

- → Fit a topic model to LDP platforms
- → Extract two topics that look like 'pork' and 'policy'
- → Average these per year and plot
- → Compare relative prevalence to electoral change timeline

EXPLAINING TOPIC PREVALENCE

Often we want to both *measure* but also *explain* the prevalence of topic mentions

Example: What are the effects of a Japanese house electoral reform on candidate platforms? (Catalinac, 2018)

- → Fit a topic model to LDP platforms
- → Extract two topics that look like 'pork' and 'policy'
- → Average these per year and plot
- → Compare relative prevalence to electoral change timeline

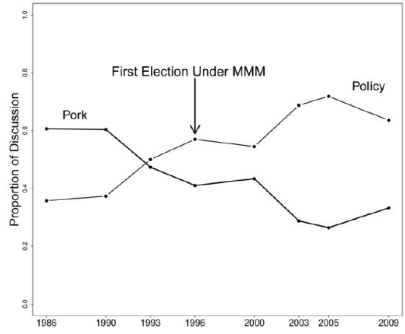


Figure 1. LDP candidates switched to more policy and less pork in the 1993 election and continued with this strategy under MMM. This figure plots the mean proportions of discussion devoted to pork and policy, respectively, in the 2,355 manifestos produced by LDP candidates in these eight elections.

STRUCTURAL TOPIC MODEL

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?

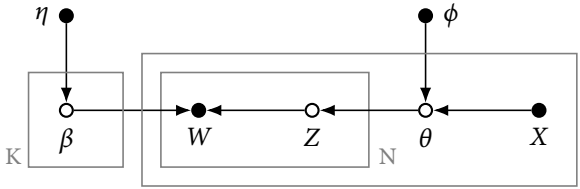
STRUCTURAL TOPIC MODEL

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 (Roberts et al., 2014) with

→ speaker party indicator, convariates etc. as X (observed)
→ proportion of the speech assigned to topic 3 as θ_3 (inferred, not observed)
→ The words W (observed)



EVEN MORE TOPIC MODELS

There's a small industry developing new types of topic model

→ A brief search will acquaint you with more than enough to play with

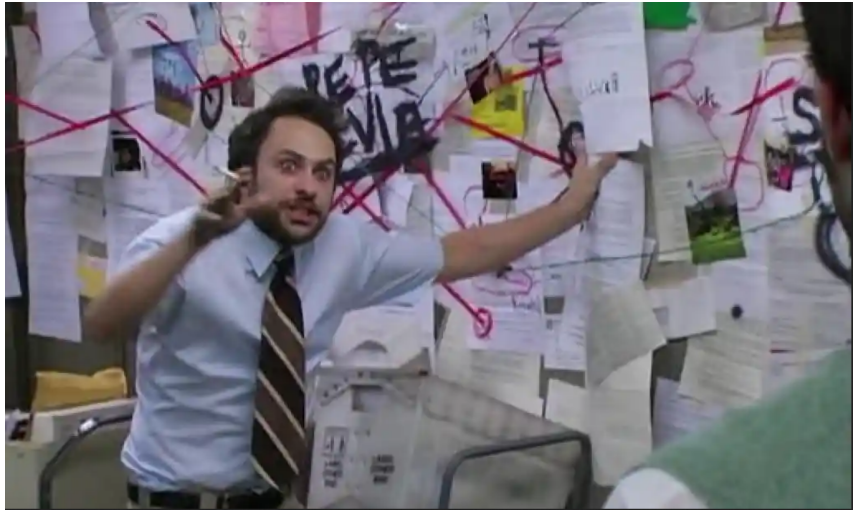
Check if they have stable code!

TEXT AS DATA

Zoom fatigue version

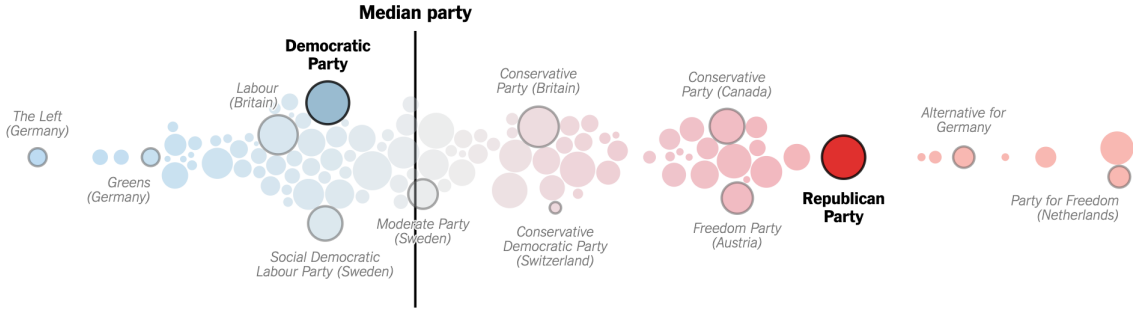
- → Topic models assume each document contains a mix of different topics
- → It attempts to infer *both* the proportion of each topic per document and the topic-word relationship (or 'dictionary')
- → *Structural topic models* allow the proportion of each topic to depend on features of each document
- → If the topic-word relationship is *known* we get 'dictionary-based content analysis'

SCALING DOCUMENTS



PROJECTION?

"what would you say if you saw this in another country?" (Brendan Nyhan)



Note: Circles sized by the percentage of the vote won by the party in the latest election in this data. Only parties that won more than 1 percent of the vote and are still in existence are shown. We analyzed parties in a selection of Western European countries, Canada and the United States.

New York Times 26.06.2019

SCALING

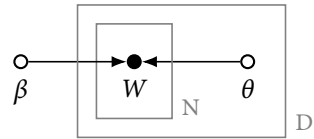
Often it's useful to think of documents living in a space

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

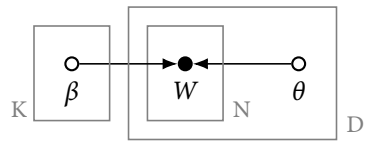
But we can also collapse them down into a *smaller space*, e.g. to 1 or K dimensions: θ

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

All we have is a term document matrix W (and assumptions)



One dimensional scaling



K-dimensional scaling

WHERE POSITIONAL INFORMATION LIVES

		Word			
	Party	Wirtschaft	soziale	Förderung	...
2002	FDP	14	4	15	
	CDU	11	8	20	
	SPD	15	9	10	
	PDS	7	16	9	
	Grüne	2	41	12	
	...				

Assumptions:

- → Position does not depend on *document length*
- → Position does not depend on *word frequency*

WHERE POSITIONAL INFORMATION LIVES

		Word			
	Party	Wirtschaft	soziale	Förderung	...
2002	FDP	14	4	15	
	CDU	11	8	20	
	SPD	15	9	10	
	PDS	7	16	9	
	Grüne	2	41	12	
	...				

Assumptions:

- → Position does not depend on *document length*
- → Position does not depend on *word frequency*

Implication

- → table margins are uninformative

WHERE POSITIONAL INFORMATION LIVES

		Word			
	Party	Wirtschaft	soziale	Förderung	…
2002	FDP	14	4	15	
	CDU	**11**	**8**	20	
	SPD	15	9	10	
	PDS	7	16	9	
	Grüne	2	41	12	
	…				

That leaves only *association structure*.

WHERE POSITIONAL INFORMATION LIVES

		Word			
	Party	Wirtschaft	soziale	Förderung	…
2002	FDP	14	4	15	
	CDU	**11**	**8**	20	
	SPD	15	9	10	
	PDS	7	16	9	
	Grüne	2	41	12	
	…				

That leaves only *association structure*.

The CDU uses 'Wirtschaft' (business) 11/8 = 1.38 times more than 'soziale' (social).

WHERE POSITIONAL INFORMATION LIVES

		Word			
	Party	Wirtschaft	soziale	Förderung	…
2002	FDP	**14**	**4**	15	
	CDU	11	8	20	
	SPD	15	9	10	
	PDS	7	16	9	
	Grüne	2	41	12	
	…				

The FDP uses 'Wirtschaft' (business) 14/4 = 3.5 times more than 'soziale' (social).

WHERE POSITIONAL INFORMATION LIVES

Many $(N-1)(V-1)$ small but relevant facts about relative proportional emphasis

1. FDP's emphasis on 'Wirtschaft' over 'soziale' is 3.5/1.375 = 2.55 times larger than that of the CDU.
2. CDU's emphasis on 'Wirtschaft' over 'soziale' is 0.82...
3. ...

You might recognize 2.55 and 0.82 and so on as *odds ratios*

$$\frac{P(\text{Wirtschaft} \mid \text{FDP})}{P(\text{soziale} \mid \text{FDP})} \Big/ \frac{P(\text{Wirtschaft} \mid \text{CDU})}{P(\text{soziale} \mid \text{CDU})} = \frac{14}{4} \Big/ \frac{11}{8}$$

which are delightfully indifferent to document lengths and word frequencies.[1]

[1] Add k the frequency of 'Wirtschaft', keeping the odds ratio the same, and notice that it just adds (some function of) k to both numerator and denominator, which cancel.

WHERE POSITIONAL INFORMATION LIVES

Actually this is where *all* substantively interesting information in document term matrices lives

→ where else is there?

Any kind of text model, e.g. a topic model

→ implies constraints on how these odds ratios can vary

→ reduces the dimensionality of word distributions to a lower than V space

So let's think about building a model of them from first principles

MODELING THE ASSOCIATIONS

First we'll assume that each C_{ij} is Poisson distributed with some rate $\mu_{ij} = \mathrm{E}[C_{ij}]$

$$C_{ij} \sim \text{Poisson}(\mu_{ij})$$

MODELING THE ASSOCIATIONS

First we'll assume that each C_{ij} is Poisson distributed with some rate $\mu_{ij} = \mathrm{E}[C_{ij}]$

$$C_{ij} \sim \text{Poisson}(\mu_{ij})$$

There are two *log-linear models* of any contingency table

$$\begin{aligned} \log \mu_{ij} &= \alpha_i + \psi_j && \text{(boring)} \\ &= \alpha_i + \psi_j + \lambda_{ij} && \text{(pointless)} \end{aligned}$$

MODELING THE ASSOCIATIONS

First we'll assume that each C_{ij} is a Poisson distributed with some expected rate

$$C_{ij} \sim \text{Poisson}(\mu_{ij})$$

There are two *log-linear models* of any contingency table

$$\begin{aligned} \log \mu_{ij} &= \alpha_i + \psi_j && \text{(independence)} \\ &= \alpha_i + \psi_j + \lambda_{ij} && \text{(saturated)} \end{aligned}$$

MODELING THE ASSOCIATIONS

First we'll assume that each C_{ij} is a Poisson distributed with some expected rate

$$C_{ij} \sim \text{Poisson}(\mu_{ij})$$

There are two *log-linear models* of any contingency table

$$\begin{aligned} \log \mu_{ij} &= \alpha_i + \psi_j && \text{(independence)} \\ &= \alpha_i + \psi_j + \lambda_{ij} && \text{(saturated)} \end{aligned}$$

All the *relative emphasis*, all the odds ratio information, and all the *position-taking* is in λ

Reminder:

→ In log linear model land, the matrix of λ values is just the same size as C

→ but the influence of the row and column margins has been *removed* by α and ψ

INFER DIMENSIONAL STRUCTURE

Intuition: λ has an orthogonal decomposition

$$\begin{aligned} \lambda &= \Theta \Sigma B^T && \text{(SVD)} \\ &= \sum_m^M \theta_{(m)} \sigma_{(m)} \beta_{(m)}^T \\ &\approx \theta\, \sigma\, \beta^T && \text{(Rank 1 approx.)} \end{aligned}$$

θ are *document positions*

β are *word positions*

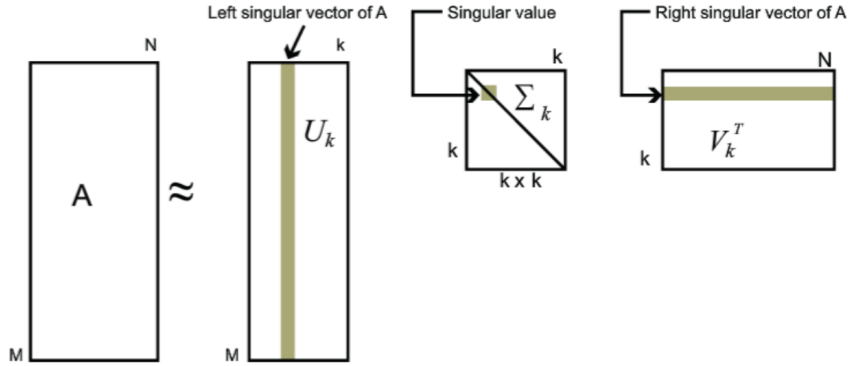
σ says *how much relative emphasizing* is happening in this dimension

So our final model is (Goodman, 1979, 1981)

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

(we'll keep the σ explicit for later)

SINGULAR VALUE DECOMPOSITION



where A is our λ, U is our θ and V is our β

In practice we'll fit it by coordinate ascent, with θ constrained to mean 0, variance 1.

THIS IS A VERY GOOD IDEA

Everybody has it...

→ Ecology, archaeology, psychology, political science

and has been having it since Hirschfeld (1935), as

→ the RC Association model (Goodman, 1981)
→ Wordfish (Slapin & Proksch, 2008)
→ Rhetorical Ideal Points (Monroe & Maeda, 2004)

THIS IS A VERY GOOD IDEA

Everybody has it...

→ Ecology, archaeology, psychology, political science

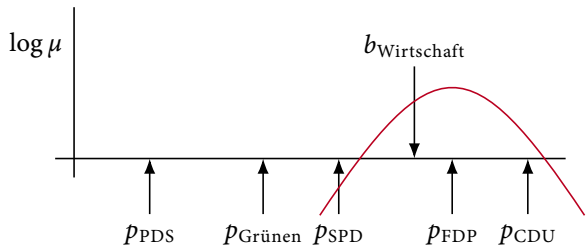
and has been having it since Hirschfeld (1935), as

→ the RC Association model (Goodman, 1981)
→ Wordfish (Slapin & Proksch, 2008)
→ Rhetorical Ideal Points (Monroe & Maeda, 2004)

That was just algebra – *why* is this a very good idea?

SPATIAL TALKING

How often will the Free Democrats (FDP) say 'Wirtschaft'?



$$\log \mu_{i,\text{Wirtschaft}} = r_i + c_{\text{Wirtschaft}} - \frac{1}{2}\frac{(p_i - b_{\text{Wirtschaft}})^2}{v}$$

$$= \underbrace{[r_i - p_i^2/v]}_{\alpha_i} + \underbrace{[c_{\text{Wirtschaft}} - b_{\text{Wirtschaft}}^2/v]}_{\psi_{\text{soziale}}} + \underbrace{p_i}_{\theta_i}\ \underbrace{1/v}_{\sigma}\ \underbrace{b_{\text{Wirtschaft}}}_{\beta_{\text{Wirtschaft}}}$$

SPATIAL TALKING

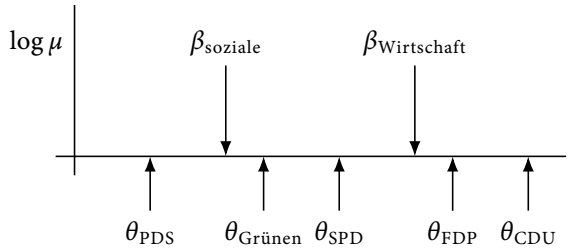
How much should the Greens say 'Wirtschaft' *or* 'soziale' in N_i words?

Condition on N_i to get a choice model (Baker, 1994; Clinton et al., 2004; Lang, 2004)

→ A multinomial logistic regression

SPATIAL TALKING

How much should the Greens say 'Wirtschaft' *or* 'soziale' in N_i words?



Condition on N_i to get a choice model (Baker, 1994; Clinton et al., 2004; Lang, 2004)

→ A multinomial logistic regression

This is a *discriminative* formulation:

$$\log\left(\frac{\pi_{i,\text{Wirtschaft}}}{\pi_{i,\text{soziale}}}\right) = \psi + \theta_i \tilde{\beta}_{\text{Wirtschaft/soziale}}$$

where $\tilde{\beta}_{\text{Wirtschaft/soziale}} = \beta_{\text{Wirtschaft}} - \beta_{\text{soziale}}$

AN IMPORTANT SPECIAL CASE

There are only *two* words, or topics, or whatever we have decided to count (Lowe et al., 2011)

$$\theta \propto \log\left(\frac{C_{i,\text{Wirtschaft}}}{C_{i,\text{soziale}}}\right)$$

AN IMPORTANT SPECIAL CASE

There are only *two* words, or topics, or whatever we have decided to count (Lowe et al., 2011)

$$\theta \propto \log\left(\frac{C_{i,\text{Wirtschaft}}}{C_{i,\text{soziale}}}\right)$$

Put another way, the model we have derived is a *generalization* of the log ratios we have been seeing previously, to

→ more than two

→ perhaps variably informative

things we can count, wrapped up as a statistical model

IF YOU CAN COUNT IT

…we can scale it

This model works for *counts*

- → all word counts
- → counts of a vocabulary subset, e.g. positive and negative affect words
- → manually assigned topic counts, e.g. a manual coding exercise
- → machine-derived topic counts , e.g. $N_i\theta_i$ (re-inflated counts) from a topic model

INTERPRETATION

What is this dimension anyway?

- → Whatever maximizes the likelihood
- → The optimal single dimensional approximation of the space of relative emphases

Substantively…we have to look

- → Which words have high and low βs?

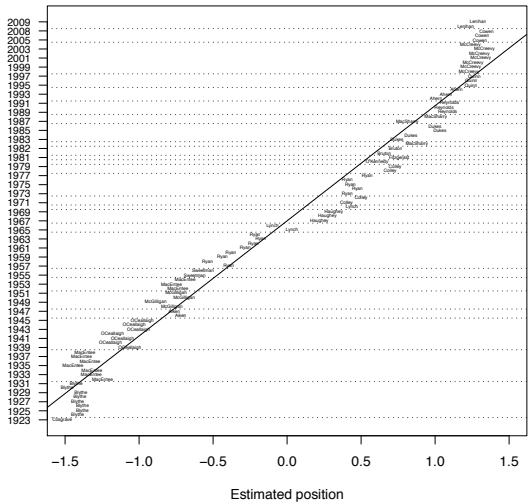
Not everything has to be a dimension

- → but it does for a scaling model!

Difficult cases:

- → Sentiment, Euroskepticism, Ethnic appeals
- → Populism and anti-system parties. Are they well understood as ideological?
- → Government and opposition. Naturally polar but not necessarily ideologically so

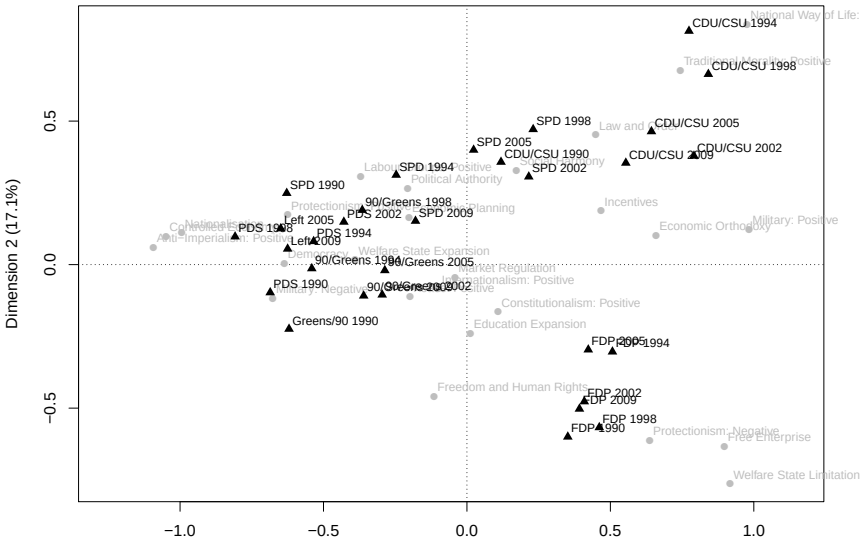
OOPS



All the budget speeches in independent Irish history, scaled. (Example courtesy of Ken Benoit)

- → Budgets are about spending money on things
- → Those things change over time
- → The model *cannot know*

PARTYING IN TWO DIMENSIONS



LIFE SKILLS

How to read a biplot:

- → Documents points are closer when using words/topics similarly
- → Words points are closer with *similar* document profiles
- → a document or word/topic used exactly as often as we would expect by chance is at 0,0
- → Document vector: arrow from 0,0 to a document point
- → Word/topic vector: arrow from 0,0 to a word/topic point
- → Vectors are *longer* the more their usage diverges from chance
- → Angle between a word vector and document vector: how much a document preferentially uses the word

TEXT AS DATA

Zoom fatigue version

- → Scaling models place documents and words in a *latent space*
- → They are the reduced form of a *spatial talking* model with quadratic utilities
- → Their induced dimensions need to be *interpreted cautiously*
- → Multiple orthogonal dimensions can also be extracted and plotted in a 'biplot'
- → *Discriminative* versions of scaling models are an open research problem

REFERENCES

Austin, J. L. (1962). "How to do things with words." Clarendon Press.

Baker, S. G. (1994). "The multinomial-Poisson transformation." *Journal of the Royal Statistical Society. Series D (The Statistician)*, *43*(4), 495–504.

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

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.

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

Clinton, J., Jackman, S., & Rivers, D. (2004). "The statistical analysis of roll call data." *American Political Science Review*, *98*(02), 1–16.

REFERENCES

Collins, P. M., Corley, P. C., & Hamner, J. (2015). "The influence of amicus curiae briefs on U.S. Supreme Court opinion." *Law & Society Review*, *49*(4), 917–944.

Corley, P. C., Collins, P. M., & Calvin, B. (2011). "Lower court influence on US Supreme Court opinion content." *The Journal of Politics*, *73*(1), 31–44.

Davidson, D. (1985). "Inquiries into truth and interpretation." Clarendon Press.

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.

Garrett, K. N., & Jansa, J. M. (2015). "Interest group influence in policy diffusion networks." *State Politics & Policy Quarterly*, *15*(3), 387–417.

REFERENCES

Giannetti, D., & Pedrazzani, A. (2016). "Rules and speeches: How parliamentary rules affect legislators' speech-making behavior: Rules and speeches." *Legislative Studies Quarterly*, *41*(3), 771–800.

Goodman, L. A. (1979). "Simple models for the analysis of association in cross-classifications having ordered categories." *Journal of the American Statistical Association*, *74*(367), 537–552.

Goodman, L. A. (1981). "Association models and canonical correlation in the analysis of cross-classifications having ordered categories." *Journal of the American Statistical Association*, *76*(374), 320–334.

Grice, P. (1993). "Studies in the way of words (3rd. printing). Harvard Univ. Press.

Grimmer, J. (2010). "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis*, *18*(1), 1–35.

Hirschfeld, H. O. (1935). "A connection between correlation and contingency." *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(4), 520–524.

REFERENCES

Jordan, M. I. (1995). *Why the logistic function?* (Computational Cognitive Science No. 9503). MIT.

Jørgensen, M., & Phillips, L. (2002). "Discourse analysis as theory and method." Sage Publications.

Klüver, H. (2009). "Measuring interest group influence using quantitative text analysis." *European Union Politics*, *10*(4), 535–549.

Lang, J. B. (2004). "Multinomial-Poisson homogeneous models for contingency tables." *The Annals of Statistics*, *32*(1), 340–383.

Laver, M., & Garry, J. (2000). "Estimating policy positions from political texts." *American Journal of Political Science*, *44*(3), 619–634.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). "The parable of Google flu: Traps in big data analysis." *Science*, *343*(6176), 1203–1205.

Lewis, D. K. (1986). "Convention: A philosophical study." Basil Blackwell. (Original work published 1969)

Lowe, W., Benoit, K. R., Mikhaylov, S., & Laver, M. (2011). "Scaling policy preferences from coded political texts." *Legislative Studies Quarterly*, *36*(1), 123–155.

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.

McCallum, A., & Nigam, K. (1993). "A comparison of event models for Naive Bayes text classification." *AAAI/ICML-98 Workshop on Learning for Text Categorization*, 41–48.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). "Optimizing semantic coherence in topic models." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272.

Monroe, B. L., Colaresi, M., & Quinn, K. M. (2008). "Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis*, *16*(4), 372–403.

Monroe, B. L., & Maeda, K. (2004). *Talk's cheap: Text-based estimation of rhetorical ideal-points*.

REFERENCES

Padua, S. (2015). “The thrilling adventures of Lovelace and Babbage: With interesting & curious anecdotes of celebrated and distinguished characters: Fully illustrating a variety of instructive and amusing scenes; as performed within and without the remarkable difference engine.” Pantheon Books.

Powell, E. (1968). “Speech delivered to the Conservative Association.”

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). “Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches.” *Legislative Studies Quarterly*, *44*(1), 97–131.

Proksch, S.-O., & Slapin, J. B. (2014). “The politics of parliamentary debate: Parties, rebels and representation.” Cambridge University Press.

Quine, W. v. O. (1960). “Word and object.” MIT Press.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2007). “How to analyze political attention with minimal assumptions and costs.” *American Journal of Political Science*, *54*(1), 209–228.

REFERENCES

Riker, W. H., Calvert, R. L., Mueller, J. E., & Wilson, R. K. (1996). "The strategy of rhetoric: Campaigning for the American constitution." Yale University Press.

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.

Searle, J. R. (1995). "The construction of social reality." Free Press.

Slapin, J. B., & Proksch, S.-O. (2008). "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science*, *52*(3), 705–722.

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., Jurgens, D., Jurafsky, D., & Eberhardt, J. L. (2017). "Language from police body camera footage shows racial disparities in officer respect." *Proceedings of the National Academy of Sciences*, *114*(25), 6521–6526.

REFERENCES

Zhang, H., & Maloney, L. T. (2012). “Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition.” *Frontiers in Neuroscience*, *6*.

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