Computer-assisted content analysis

Will Lowe Princeton University
James Lo University of Southern California

Practicalities: Materials

We have a website:

`http://ec2-52-207-214-68.compute-1.amazonaws.com:8787`

Password: iqmr2016

Labs

Menu

Session 0: How could this possibly work?

Session 1: Dictionary-based ‘classical’ content analysis

Session 2: Classification and topic models

Session 3: Scaling models

Focus

Assumptions

Mechanics (in Lab)

Interpretation

Pitfalls

Topics

How to learn about

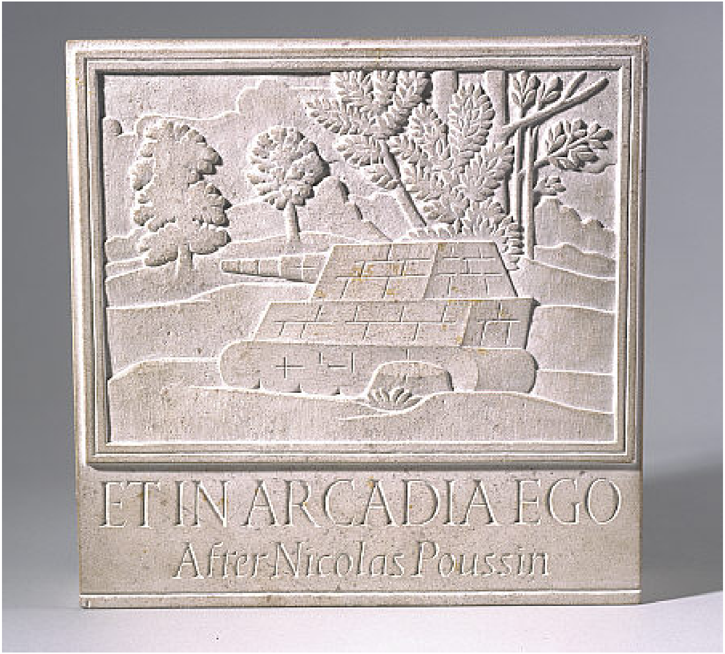
- party platforms
- legislative agendas
- parliamentary debates
- bloggers
- presidents
- international terrorists

by counting (lots of) words…

The transcendental question

What are the conditions for the possibility of learning about these things by counting words?

how could this possibly work?


ET IN ARCADIA EGO
After Nicolas Poussin

Big picture

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

> The topic of my lecture, my position on a political issue, the importance of defence issues to a some political party.

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

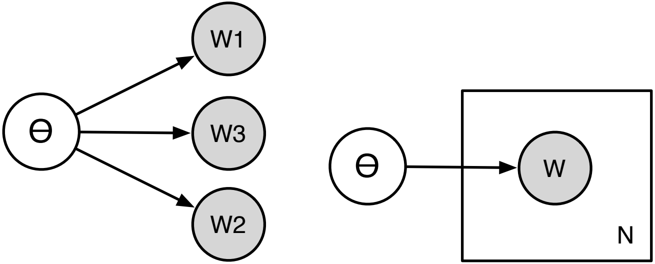
> yelling, muttering, cursing, lecturing

which *can* be directly observed.

Focus on the *expressed message* and the *words*…

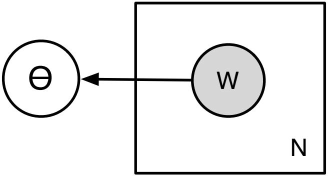
Communication

To *communicate* a message θ – to inform, persuade, demand, threaten, a producer (the speaker or writer) *generates* words of different kinds in different quantities



Communication

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



Communication

This is a stable (Searle, 1995) conventional (Lewis, 1969) but disruptable (Riker, 1996) communication process in which no finite set of words *uniquely* identifies any content (Quine, 1960; Davidson, 1977)

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

Rely on:

- instrumentality
- reflexivity
- randomness

Instrumentality from 'them'

Language use is as a *form of action* (Wittgenstein, 1953; Austin, 1975; Dawkins and Krebs, 1978)

Note the distinction between

> ‘*W* means *X*’
>
> versus
>
> ‘*W* is used to mean *X*’

Instrumentality from us

The secret of quantitative political text analysis:

we aren't actually interested in words W

that's for linguists…

we aren't actually interested in what's in your head θ

that's for psychologists…

except as they help explain things we are interested in.
They are *just data*.

Reflexivity

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

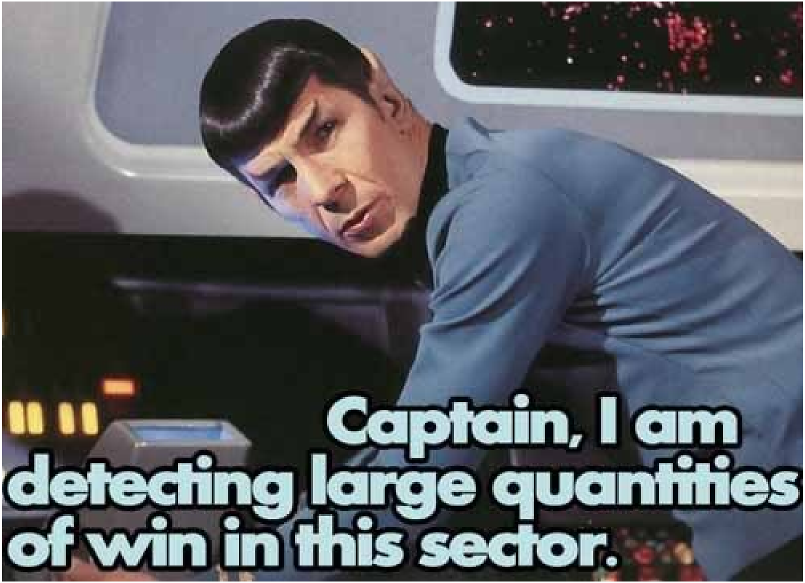
> *My theme here has, as it were, four heads. […] The first is articulated by the word "opportunity" […] the second is expressed by the word "choice" […] the third theme is summed up by the word "strength" [and] my fourth theme is expressed well by the word "renewal"*
> *(M. Thatcher, 1979)*

[2, 7, 2, 8] in 4431 words

Reflexivity

Or maybe just one theme…

> *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.

Scope conditions

Computer-assisted content analysis works best when language usage is

stable, conventionalized, and instrumental

Implicitly, we usually condition on some institution, e.g.

courts, legislatures, online political argument, sports or financial reporting, survey responses

Scope conditions

Computer-assisted content analysis works best when language usage is

stable, conventionalized, and instrumental

Implicitly, we usually condition on some institution, e.g.

courts, legislatures, online political argument, sports or financial reporting, survey responses

(Notice that this inevitably creates a comparability problem)

Randomness

You almost never *say exactly the same words twice*, even when you haven’t changed your mind about the message.

Randomness

You almost never *say exactly the same words twice*, even when you haven’t changed your mind about the message.

Hence words are the result of some kind of *sampling process*.

We treat this process as *random* because we don’t know or care about all the causes of variation

Randomness

You almost never *say exactly the same words twice*, even when you haven’t changed your mind about the message.

Hence words are the result of some kind of *sampling process*.

We treat this process as *random* because we don’t know or care about all the causes of variation

(and because we’re all secretly Bayesians)

Words as data

What do we know about words *as data*?

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: Labour party (2010) manifesto compared to other parties in two elections

High D. 8038 word types in two elections
(adult native english speakers know $\sim$20-35,000)

Sparse Of these, Labour only uses 4273 (53.16%)

Skewed Of these 1703 (21.19%) words appear exactly once, and 949 (11.81%) appear <5 times

Difficult words

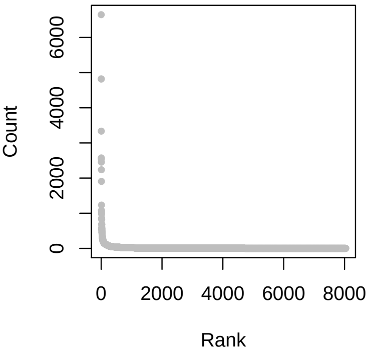
Words are not like your other data...

Zipf-Mandelbrot law (a pareto distribution in disguise)

$$P(w_i) \propto 1/r_i^a$$

where r_i is the frequency *rank* of word i and $a \approx 1$

Very fat tailed…


Count
0
2000
4000
6000
0
2000
4000
6000
8000
Rank
Log count
1
5
50
500
5000
1
10
100
1000
Log rank

Dealing with difficult words

Frequency is inversely proportional to substantive interestingness

Bottom 10:

	Count
dream	1
flair	1
world-beating	1
globally-respected	1
underdog	1
heading	1
frustrations	1
unruly	1
walk	1
out-of-control	1

Dealing with difficult words

Top 10

	Count
the	6648
and	4823
to	4817
of	3335
will	2574
we	2546
a	2454
in	2237
for	1905
that	1232

Dealing with difficult words

Top 10 minus the ‘standard’ stopwords

	Count
will	2574
people	692
new	559
government	458
local	404
work	354
support	334
britain	326
make	322
public	311

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 a 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*

The standard set of equivalence assumptions are the ‘bag of words’.

Specifically:

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'..."
(E. 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'..."
(E. 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

		unit	
		‘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
	…	…	…

Count data

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

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

Count data

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

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

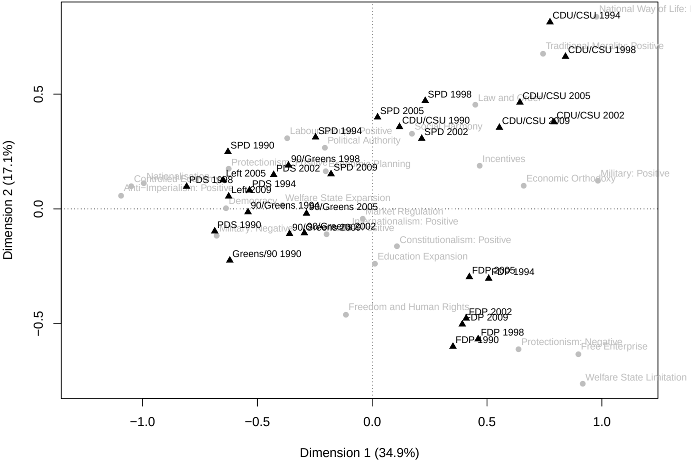
Everything you learned in your categorical data analysis course applies

except that the variables of interest: θ are *not observed*

What we want

	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}		

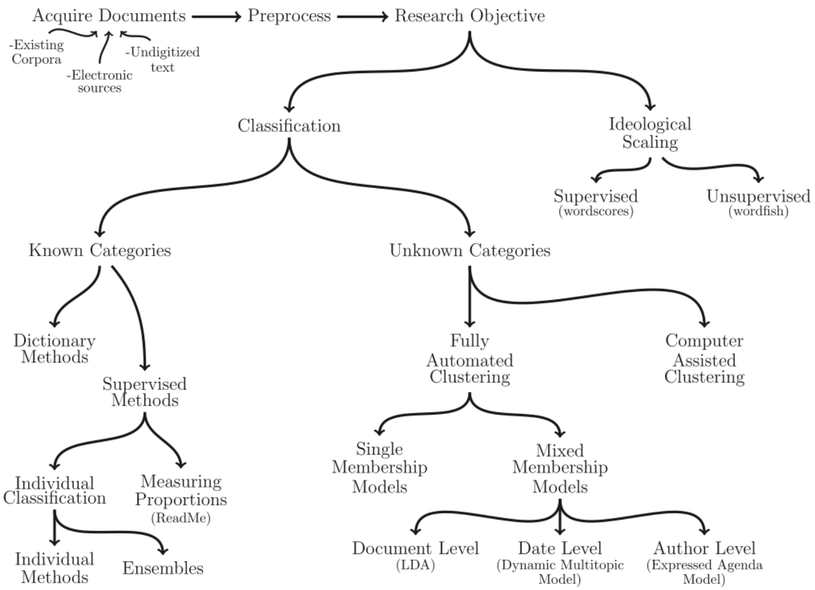
Visualized

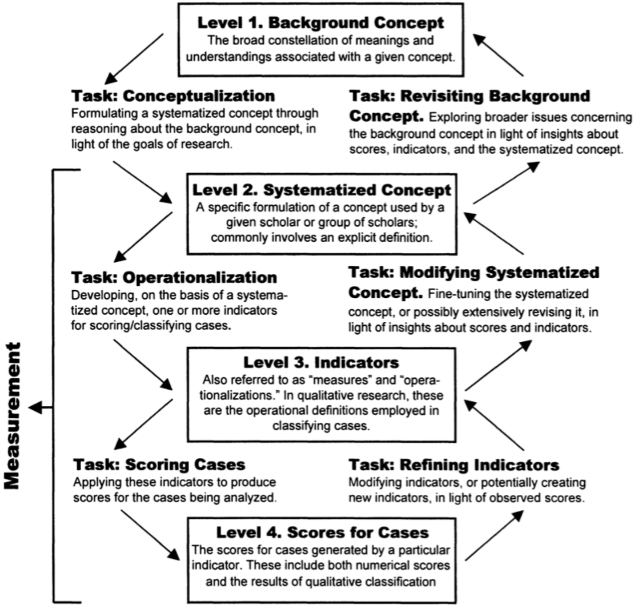

National Way of Life: I
CDU/CSU 1994
Traditional Morality: Positive
CDU/CSU 1998
SPD 1998
Law and Order
CDU/CSU 2005
SPD 2005
CDU/CSU 1990
CDU/CSU 2009
CDU/CSU 2002
Labour Groups: Positive
SPD 1994
Social Harmony
SPD 2002
Political Authority
SPD 1990
Incentives
Protectionism: Positive
90/Greens 1998
Economic Planning
PDS 2002
SPD 2009
Left 2005
Nationalisation
Controlled Economy
PDS 1998
PDS 1994
Economic Orthodoxy
Military: Positive
Anti−Imperialism: Positive
Left 2009
Democracy
Welfare State Expansion
90/Greens 1994
90/Greens 2005
Market Regulation
Internationalism: Positive
PDS 1990
Military: Negative
90/Greens 2009
90/Greens 2002
Constitutionalism: Positive
Education Expansion
Greens/90 1990
FDP 2005
FDP 1994
Freedom and Human Rights
FDP 2002
FDP 2009
FDP 1998
FDP 1990
Protectionism: Negative
Free Enterprise
Welfare State Limitation
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
Dimension 2 (17.1%)
Dimension 1 (34.9%)

What is this content θ?

What is the content in content analysis?

- Documents are *mixtures of categories*: policy agenda of a speech
- Documents have *categories*: topic of a press release
- Documents have *positions*: ideological position of a legal brief


Acquire Documents
Preprocess
Research Objective
-Existing Corpora
-Electronic sources
-Undigitized text
Classification
Ideological Scaling
Supervised (wordscores)
Unsupervised (wordfish)
Known Categories
Unknown Categories
Dictionary Methods
Supervised Methods
Fully Automated Clustering
Computer Assisted Clustering
Individual Classification
Measuring Proportions (ReadMe)
Single Membership Models
Mixed Membership Models
Individual Methods
Ensembles
Document Level (LDA)
Date Level (Dynamic Multitopic Model)
Author Level (Expressed Agenda Model)



Level 1. Background Concept
The broad constellation of meanings and understandings associated with a given concept.

Task: Conceptualization
Formulating a systematized concept through reasoning about the background concept, in light of the goals of research.

Task: Revisiting Background Concept. Exploring broader issues concerning the background concept in light of insights about scores, indicators, and the systematized concept.

Level 2. Systematized Concept
A specific formulation of a concept used by a given scholar or group of scholars; commonly involves an explicit definition.

Task: Operationalization
Developing, on the basis of a systematized concept, one or more indicators for scoring/classifying cases.

Task: Modifying Systematized Concept. Fine-tuning the systematized concept, or possibly extensively revising it, in light of insights about scores and indicators.

Level 3. Indicators
Also referred to as "measures" and "operationalizations." In qualitative research, these are the operational definitions employed in classifying cases.

Task: Scoring Cases
Applying these indicators to produce scores for the cases being analyzed.

Task: Refining Indicators
Modifying indicators, or potentially creating new indicators, in light of observed scores.

Level 4. Scores for Cases
The scores for cases generated by a particular indicator. These include both numerical scores and the results of qualitative classification

Commitment issues

What are we committing to in this quantitative content analysis framework?

Probably less than you think...

Assumptions:

- θ is socially/institutionally constructed: only linguists care about the real thing
- There are no differences in θ that make no verbal difference (basically Pragmatism)

Theory / measurement separation

Discourse analytic approaches tend to *tightly couple* theory and 'measurement' components

(This is contingent…)

We will try as far as possible to separate them…

Our concerns: validity, stability

Rely on: transparency, reliability, replicability

Statistical models of words: Poisson

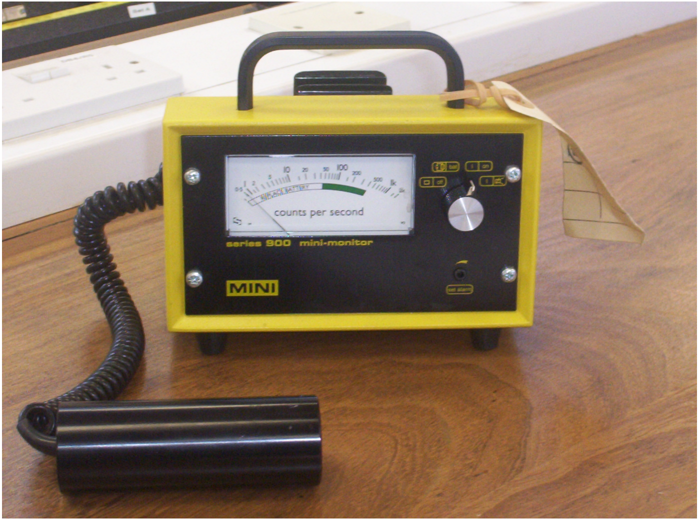
Word counts/rates are conditionally Poisson:

$$W_j \sim \text{Poisson}(\lambda_j)$$

Expected W_j (and its variance) is λ_j

Models are naturally *multiplicative*. Rates increase by 10%, decrease by 20%

Conditional on what? Typically on θ


0.5
1
2
5
10
20
50
100
200
500
1k
2k
REPLACE BATTERY
counts per second
series 900 mini-monitor
MINI
bat
on
off
set alarm

Statistical models of words: Multinomial

For fixed document lengths, counts are conditionally Multinomial:

$$W_1 \ldots W_V \sim \text{Multinomial}(W_1 \ldots W_V; \pi_1 \ldots \pi_V, N_i)$$

Expected W_i is $N\pi_i$

Covariance of W_i and W_j is $-N\pi_i\pi_j$ (budget constraint)

18
14
16
20
11
7
3

Implication: Absence is an observation

Don’t be fooled…

Statistical models of text deal with *absence* as well as presence: zeros count

Absence is informative *to the extent it is surprising*

Surprise implies expectations; expectations imply a model.

Looking ahead: Modeling strategies

We can model the content of a term-document matrix in several ways

$\theta \longleftarrow$ words: Go for $P(\theta \mid \text{words})$ *directly*

Requires some *observed* θ, and lots of *careful* regression modeling, or manual coding

$\theta \longrightarrow$ words: Get $P(\theta \mid \text{words})$ *indirectly*

Model words as a function of θ, add a prior, and infer θ using Bayes theorem

$$P(\theta \mid \text{words}) = \frac{P(\text{words} \mid \theta)P(\theta)}{\sum_k^{\theta} P(\text{words} \mid \theta_k)P(\theta_k)}$$

Classical content analysis

Content is, or is constructed from, *categories* e.g.

> human rights, welfare state, national security

Substantively these often have *valence*, e.g.

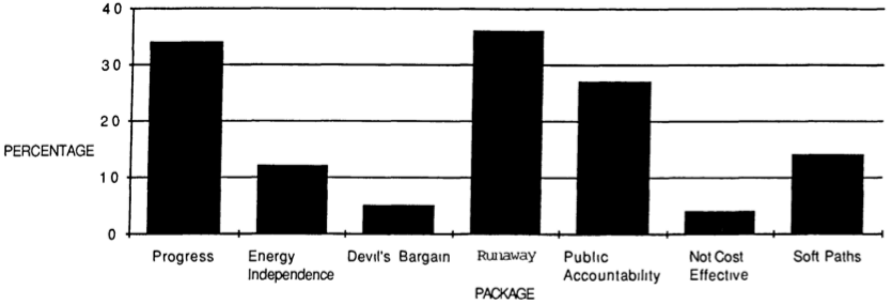
> pro-welfare state vs. anti-welfare state, lots of CMP categories

But they are invariably treated as *nominal level* variables

We are typically interested in them for

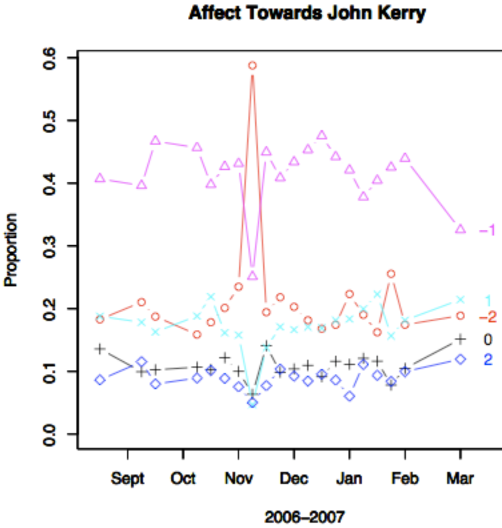
> simple descriptions, making comparisons, tracing temporal dynamics

Talking Like a newspaper

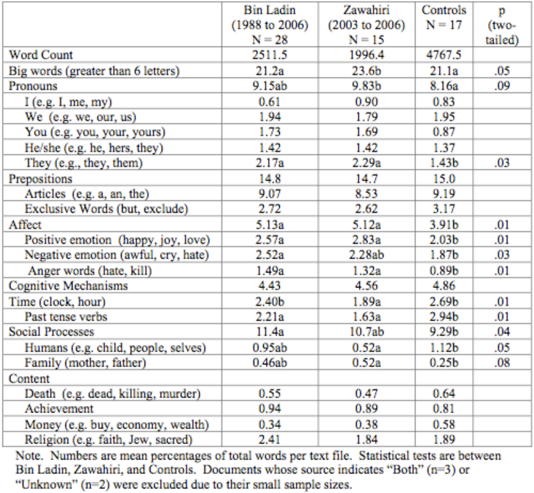


Gamson and Modigliani (1989)

Talking like a candidate



Talking like a terrorist

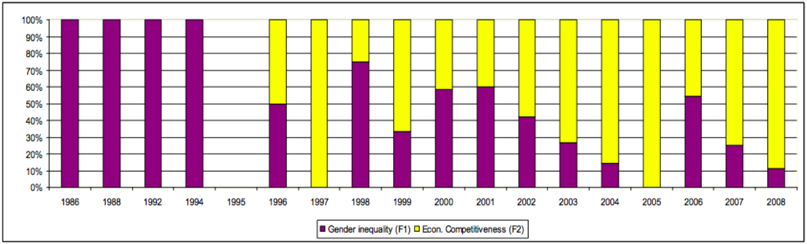


	Bin Ladin (1988 to 2006) N = 28	Zawahiri (2003 to 2006) N = 15	Controls N = 17	p (two-tailed)
Word Count	2511.5	1996.4	4767.5	
Big words (greater than 6 letters)	21.2a	23.6b	21.1a	.05
Pronouns	9.15ab	9.83b	8.16a	.09
I (e.g. I, me, my)	0.61	0.90	0.83	
We (e.g. we, our, us)	1.94	1.79	1.95	
You (e.g. you, your, yours)	1.73	1.69	0.87	
He/she (e.g. he, hers, they)	1.42	1.42	1.37	
They (e.g., they, them)	2.17a	2.29a	1.43b	.03
Prepositions	14.8	14.7	15.0	
Articles (e.g. a, an, the)	9.07	8.53	9.19	
Exclusive Words (but, exclude)	2.72	2.62	3.17	
Affect	5.13a	5.12a	3.91b	.01
Positive emotion (happy, joy, love)	2.57a	2.83a	2.03b	.01
Negative emotion (awful, cry, hate)	2.52a	2.28ab	1.87b	.03
Anger words (hate, kill)	1.49a	1.32a	0.89b	.01
Cognitive Mechanisms	4.43	4.56	4.86	
Time (clock, hour)	2.40b	1.89a	2.69b	.01
Past tense verbs	2.21a	1.63a	2.94b	.01
Social Processes	11.4a	10.7ab	9.29b	.04
Humans (e.g. child, people, selves)	0.95ab	0.52a	1.12b	.05
Family (mother, father)	0.46ab	0.52a	0.25b	.08
Content				
Death (e.g. dead, killing, murder)	0.55	0.47	0.64	
Achievement	0.94	0.89	0.81	
Money (e.g. buy, economy, wealth)	0.34	0.38	0.58	
Religion (e.g. faith, Jew, sacred)	2.41	1.84	1.89	

Note. Numbers are mean percentages of total words per text file. Statistical tests are between Bin Ladin, Zawahiri, and Controls. Documents whose source indicates "Both" (n=3) or "Unknown" (n=2) were excluded due to their small sample sizes.

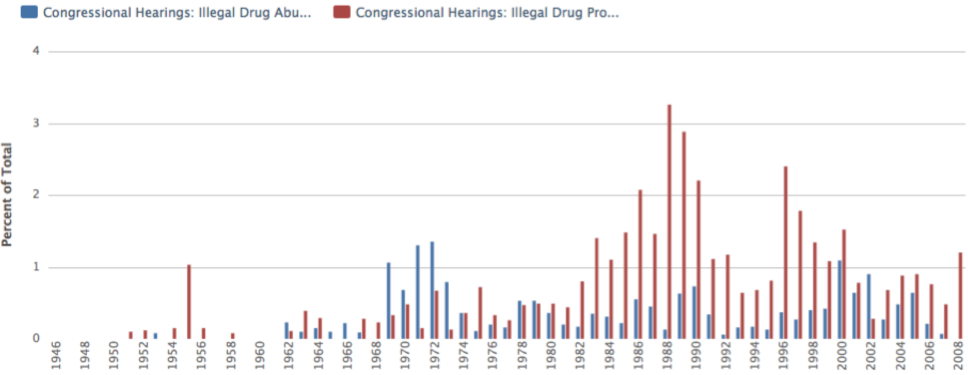
Talking like the European Commission

Figure 4.2-2 Relative proportions of policy frames F1 and F2 in secondary EU legislation



Source: Radulova (2009)

Talking About drugs



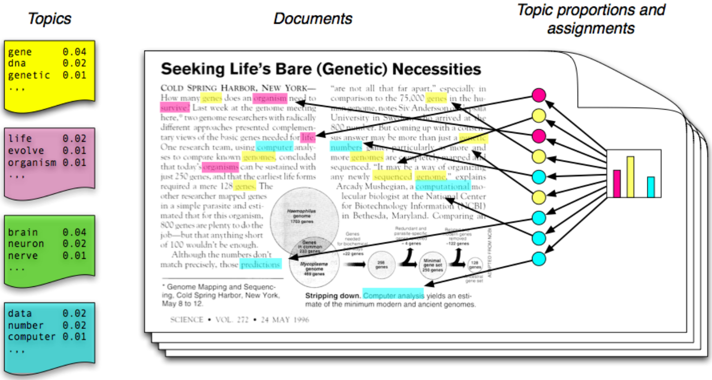
The Congressional Bills Project website (retrieved 2010)

Classical content analysis

Categories are

equivalence classes over words

representable as assignments of a K-valued category membership variable Z to each word



Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

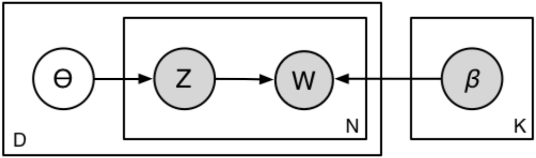
SCIENCE • VOL. 272 • 24 MAY 1996

Classical content analysis

Every word *W* has an topic *Z*

The word *W* to topic *Z* mapping β is provided by the researcher as a *content analysis dictionary*

The content of a document θ is the proportion (or count) of each category



How content is generated and what we claim to know

Content analysis dictionary

ECONOMY +STATE
accommodation
age
ambulance
assist
benefit
…
-STATE
assets
bid
choice*
compet*
constrain*
…

from Laver and Garry’s (2000) dictionary

As a posterior: P(Z | W)

Dictionary is an explicit and very *certain* statement of $P(Z \mid W)$

	Z	state reg	market econ
W	age	1	0
	benefit	1	0
	…	…	…
	assets	0	1
	bid	0	1
	…	…	…

…from a underspecified likelihood

The *only* way this could be true is if the data had been generated like

$P(W \mid Z)$

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

…leading to a posterior over content

Define the category *counts*

$$Z_k = \sum_i^N P(Z = k \mid W_i)$$

and estimate category relative *proportions* using

$$\hat{\theta}_k = \frac{Z_k}{\sum_j^K Z_j}$$

(When θ is a set of multinomial parameters, *and the model assumptions are correct*, this could be a reasonable estimator)

Reconstruction

Dictionary-based content analysis was *not* developed this way

Originally (e.g. Stone 1966) there was no probability model at all

Connecting CCA 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?

Inference About content

Statistically speaking, the three types of measures are

a proportion

a difference of proportions

a ratio of proportions

Under certain sampling assumptions we can make inferences about a population

Inference About proportions

Example: in the 2001 Labour manifesto there are 872 matches to Laver and Garry's *state reg* category

0.029 (nearly 3%) of the document's words

0.066 (about 6%) of words that matched *any* categories

The document has 30157 words, so the *first* proportion is estimated as

$$\hat{\theta}_{\textit{state reg}} \;=\; 0.029 \;\; [0.027, 0.030]$$

What does this mean?

Inference about proportions

Think of the party headquarters repeatedly *drafting* this manifesto

The true proportion – the one suitable to the party's policies – is fixed but every draft is slightly different

The confidence interval reflects the fact that we expect long manifestos to have more precise information about policy

This interval is computed as if every word was a new (conditionally) independent piece of of information

Reporting: Rates

Don’t report proportions if you don’t need to.

Rates/ratios are more intuitive

e.g. the rate of dictionary matches per B words is

$$\lambda_B = \theta B$$

which is a more interpretable proportion, e.g.

> 29 times per 1000 words

Different measures correspond to different choices of B.

Ratios: How new was New Labour?

Was the Conservative party in 1992 more or less for state intervention than 'New' Labour in 1997?

Compare instances of *state reg* and *market econ* in the manifestos

Party	Counts	
	state reg	*market econ*
Conservative	320	643
Labour	396	268

Risk ratios

Compute two *risk ratios*:

$$\begin{aligned} RR_{\mathit{state\ reg}} &= \frac{P(\mathit{state\ reg} \mid \text{cons})}{P(\mathit{state\ reg} \mid \text{lab})} \\ RR_{\mathit{market\ econ}} &= \frac{P(\mathit{market\ econ} \mid \text{cons})}{P(\mathit{market\ econ} \mid \text{lab})} \end{aligned}$$

and 95% confidence intervals

Interpreting risk ratios

If $RR = 1$ then the category occurs at the same rate in labour and conservative manifestos

If $RR = 2$ then the conservative manifesto contains *twice* as much *state reg* language as the labour manifesto

If $RR = .5$ then the conservative manifesto contains *half* as much *state reg* language as the labour manifesto

If the confidence interval for *RR* contains 1 then we *no evidence* that *state reg* and *market econ* occur at different rates

Risk ratios

	Risk Ratio
market econ	1.45 [1.26, 1.67]
state reg	0.49 [0.42, 0.57]

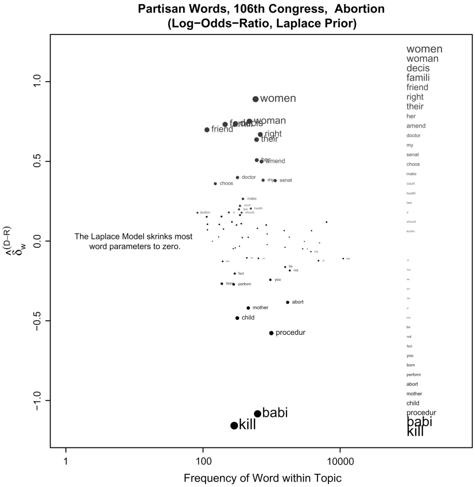
Conservative manifesto generates *market econ* words 45% more often

45% = 100(1.45 - 1)%

Conservative manifesto only generates 49% as many *state reg* words as Labour.

Equivalently Labour generates them about *twice* as often

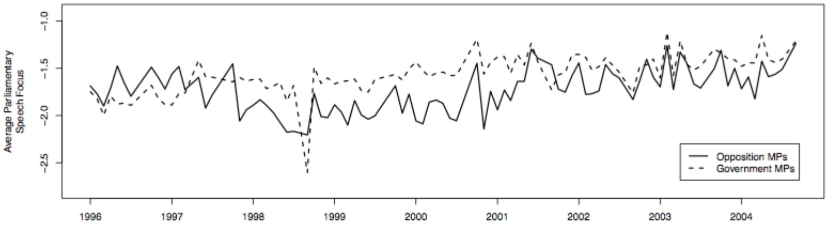
(Regularised) log ratios



The Laplace Model skrinks most word parameters to zero.

…as dependent variable

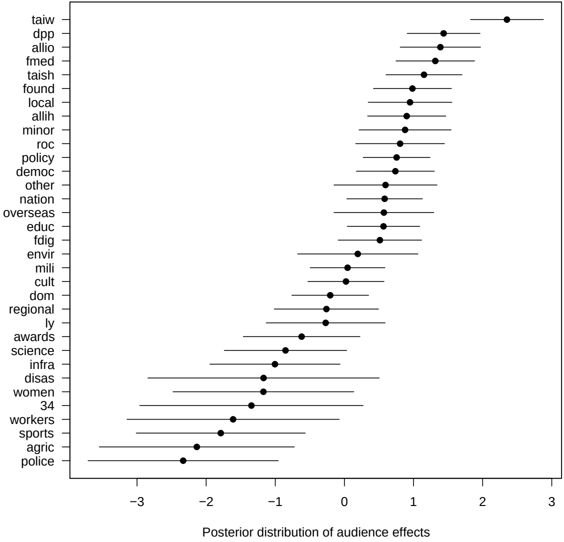
Example: district vs party focus



Data: [*district words*, *party words*] (Kellerman & Proksch, MS)

Here, a *logged ratio* of two categories

Content as something to explain



OK, how do I make such a dictionary?

Find a suitable tool

Maximise measurement validity

Minimise *measurement error*

OK, how do I make such a dictionary?

Find a suitable tool

Maximise measurement validity

Minimise *measurement error*

(Sell high, buy low)

Find a suitable tool

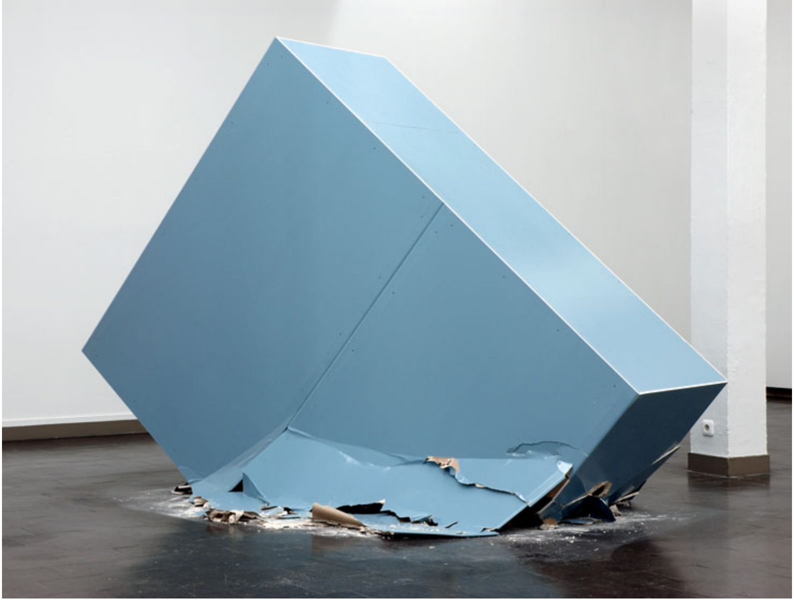
Wordstat

LIWC (maybe don’t)

Hamlet

Atlas-ti (?)

Yoshikoder



The source of measurement error

Measurement error in classical content analysis is primarily failure of *this* assumption:

	$P(Z = \textit{state reg} \mid \text{W})$	$P(Z = \textit{market econ} \mid \text{W})$
age	1	0
benefit	1	0
…	…	…
assets	0	1
bid	0	1
…	…	…

Consequences of measurement error

What are the effects of measurement error in category counts?

- Being directly wrong, e.g.

  Estimated rates are too *low* (bias)

  Some of estimates are more biased than others

- Being *indirectly* wrong, e.g.

  Subtractive or ratio left-right measures are too *centrist*

Measurement error: example

Assume

a vocabulary of only two words ‘benefit’ and ‘assets’

a *subtractive* measure of position: $Z_{market\ econ} - Z_{state\ reg}$

Then we hope that

	$P(Z = state\ reg \mid \mathrm{W})$	$P(Z = market\ econ \mid \mathrm{W})$
benefit	1	0
assets	0	1

Measurement error: example

but what if…

	state reg	*market econ*
$P(\text{benefit} \mid Z)$	0.7	0.2
$P(\text{assets} \mid Z)$	0.3	0.8

P(W='asset' | Z=*state reg*) > 0

so

P(Z=*state reg* | W='asset') < 1

Measurement error: example

Assume

$$Z_{market\ econ} = 10$$

$$Z_{state\ reg} = 20$$

Then the *true* difference is

$$\frac{(10 - 20)}{(10 + 20)} = -0.33 \tag{1}$$

Under perfect measurement this would be realised on average as

20 ‘benefit’

10 ‘assets’

Measurement error: example

Under *imperfect* measurement it is realised on average as

16 ‘benefit’
(14 from *state reg* but 2 from *market econ*)

14 ‘assets’
(8 from *market econ* but 6 from *state reg*)

Measurement error: example

The proportional difference measure is now

$$\frac{(14-16)}{(14+16)} = -0.07 \tag{2}$$

Apparently much closer to the centre, but only because of measurement error

Measurement error: example

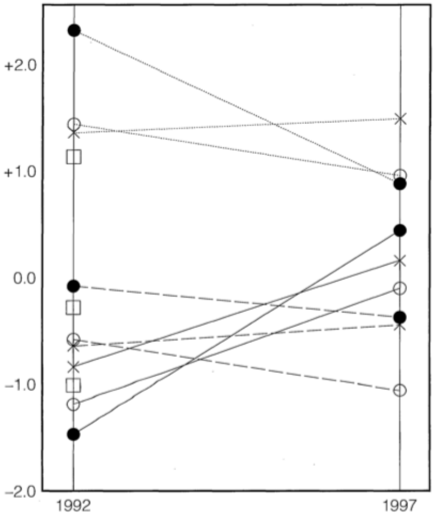
The proportional difference measure is now

$$\frac{(14-16)}{(14+16)} = -0.07 \tag{2}$$

Apparently much closer to the centre, but only because of measurement error

All relative measures will have this problem

In action (Laver and Garry 2000)



In action with people, not dictionaries

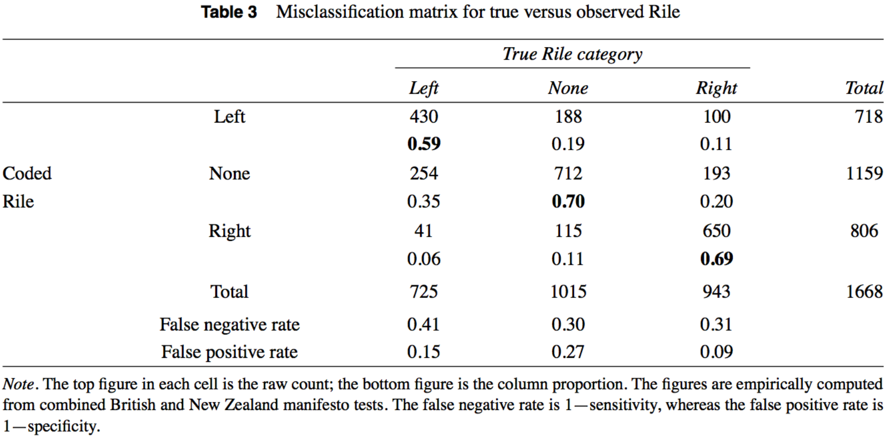
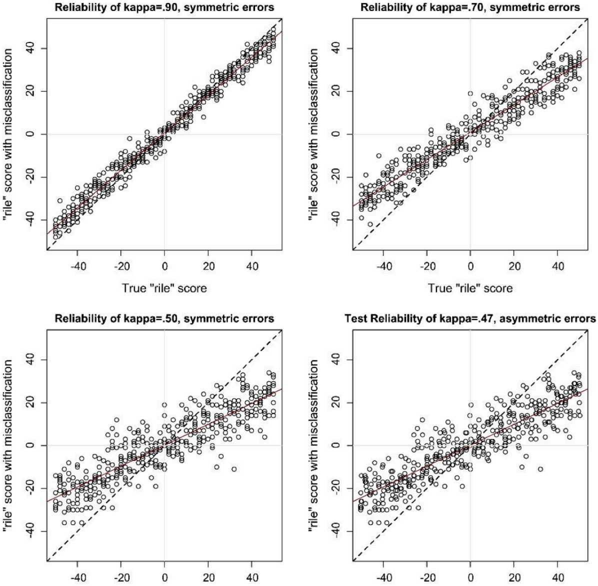


Table 3 Misclassification matrix for true versus observed Rile

		True Rile category			
		Left	*None*	*Right*	*Total*
	Left	430	188	100	718
		0.59	0.19	0.11	
Coded	None	254	712	193	1159
Rile		0.35	**0.70**	0.20	
	Right	41	115	650	806
		0.06	0.11	**0.69**	
	Total	725	1015	943	1668
	False negative rate	0.41	0.30	0.31	
	False positive rate	0.15	0.27	0.09	

Note. The top figure in each cell is the raw count; the bottom figure is the column proportion. The figures are empirically computed from combined British and New Zealand manifesto tests. The false negative rate is 1—sensitivity, whereas the false positive rate is 1—specificity.

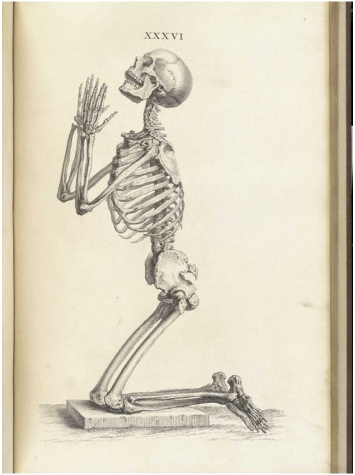
Attentuation (Mikhaylov et al. 2012)



Solutions: A quasi-theological approach

Solutions: A quasi-theological approach

‘Thoughts and prayers’



Solutions: avoid it

An often non-obvious fact about content dictionaries:

precision: proportion of words used the way your dictionary assumes

recall: proportion of words used that way that are in your dictionary

always trade-off…

Aside: precision and recall

Every field reinvents this distinction:

precision and recall

specificity and sensitivity

users and producer’s accuracy

type 1 and type 2 error

sins of omission and sins of commission

Tools to evaluate items

Keyword in context analyses (KWIC) allow you to scan all contexts of a word

- How many of them are the sense or usage you want?

	contextPre	keyword	contextPost
1	also keep all the other	benefits	that pensioners currently receive,
2	regulation will have to have	benefits	exceeding costs, and regulations
3	and Controlled Immigration Britain has	benefited	from immigration. We all
4	positive contribution But if those	benefits	are to continue to flow
5	Nor ther n Ireland brings	benefits	to all parts of our
6	their home, will also	benefit	first- time buyers.
7	you help yourself; you	benefit	and the country benefits.
8	you benefit and the country	benefits	. So now, I
9	result of our tax and	benefit	measures compared to 1997.
10	result of personal tax and	benefit	measures introduced since 1997,
11	, the savings on unemployment	benefits	will go towards investing more
12	trebled the number on incapacity	benefits	. We will help 17
13	Work programme and reform Incapacity	Benefit	, with the main elements
14	main elements of the new	benefit	regime in place from 2008
15	stronger penalties. To the	benefit	of business and household consumers
16	effective directive to provide real	benefits	to consumers and new opportunities
17	better.We are examining the potential	benefits	of a parallel Expressway on
18	ways to lock in the	benefit	of new capacity. We
19	are determined to spread the	benefits	of enterprise to every community
20	to get ahead, to	benefit	from improving public services,
21	of the school workforce is	benefiting	staff and helping to tailor
22	teachers and pupils get the	benefit	of the range of support

Measurement error: Confession and forgiveness

Under measurement error

A observed category proportions are generated by a *mixture* of categories

The weights for this mixture are the true category proportions

Given the error matrix, we can *infer* the true proportions

Intuition

$$P(W) = \sum_{k}^{K} P(W \mid Z = k) P(Z = k)$$

has the form

$$Y = X\theta$$

Measurement error: model it

In our previous example

$$\begin{bmatrix} 0.53 \\ 0.46 \end{bmatrix} = \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

is solved exactly as [0.66, 0.33]

by *inverting* the error matrix

Measurement error: model it

Applied to Mikhaylov human error data:

	L	N	R
L	430	188	100
N	254	712	193
R	41	115	650

$\Longrightarrow$

	L	N	R
L	0.59	0.19	0.11
N	0.35	0.70	0.20
R	0.06	0.11	0.69

Implication:

If [L, N, R] were [20, 0, 10] we would *expect* to see about [13, 9, 8]

Measurement error: model it

Invert $P(C \mid T)$:

	L	N	R
L	2.00	-0.50	-0.16
N	-1.00	1.75	-0.37
R	0.00	-0.25	1.52

and multiply to get an estimate of the true counts…

Example:

$$[13, 9, 8] \longrightarrow [20.19, -0.16, 9.98] \approx [20, 0, 10]$$

Notes:

Some patterns of measurement error cannot be corrected for…

These results hold *in expectation*.

We are ignoring measurement error *in the error matrix*

This is a linear method that may violate prior constraints

Works for *anything that makes errors* (human or machine)

Up next

Topic models, e.g. Latent Dirichlet Allocation (Blei et al.) we

> Build this idea into a complete model
>
> *infer* rather than assert the relationship between W and Z by learning β.

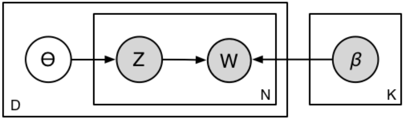
Up next

Topic models, e.g. Latent Dirichlet Allocation (Blei et al.) we

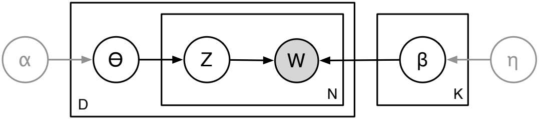
Build this idea into a complete model

infer rather than assert the relationship between W and Z by learning β.

From



to



Lab time