Day 4: Dictionary-Based Coding

Kenneth Benoit

CEU April 14-21 2011

April 19, 2011

Rationale for dictionaries

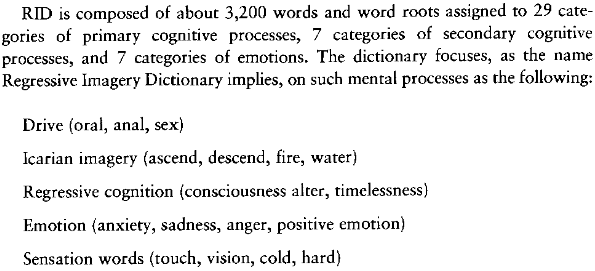
- Rather than count words that occur, pre-define words associated with specific meanings

- Frequently involves lemmatization: transformation of all inflected word forms to their “dictionary look-up form” — more powerful than stemming

- Example: General Inquirer codes *I*, *me*, *my*, *mine*, *myself* as self, and *we*, *us*, *our*, *ours*, *ourselves* as selves

Well-known dictionaries

- General Inquirer (Stone et al 1966)
- Linquistic Inquiry and Word Count (LIWC – Penaker et al 2001)
- Regressive Imagery Dictionary (Martindale 1990)



RID is composed of about 3,200 words and word roots assigned to 29 categories of primary cognitive processes, 7 categories of secondary cognitive processes, and 7 categories of emotions. The dictionary focuses, as the name Regressive Imagery Dictionary implies, on such mental processes as the following:

Drive (oral, anal, sex)

Icarian imagery (ascend, descend, fire, water)

Regressive cognition (consciousness alter, timelessness)

Emotion (anxiety, sadness, anger, positive emotion)

Sensation words (touch, vision, cold, hard)

Content analysis dictionary

```
ECONOMY / +STATE
    accommodation
    age
    ambulance
    assist
    ...

ECONOMY / -STATE
    choice*
    compet*
    constrain*
    ...
```

from Laver and Garry (2000) dictionary

As Measurement

Translation. For each word:

	$P(\theta = \text{'Pro-State'} \mid \text{W})$	$P(\theta = \text{'Anti-State'} \mid \text{W})$
age	1	0
benefit	1	0
...	...	...
assets	0	1
bid	0	1
...	...	...

Using a dictionary

For each word W_i in a document

- If W_i is in category j, increment C_j
- Compute category proportions:

$$\hat{\theta}_i = \frac{C_i}{\sum_j C_j}$$

- The vector of category proportions is the content

Using a dictionary

A wrinkle in the interpretation: No category $K+1$ to catch boring words —

> θ_i *is the proportion of category* i*, relative to other categories*

There is a category $K+1$ to catch boring words —

> θ_i *is the proportion of the document devoted to category* i

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

The large sample standard error for the proportion $\hat{\theta}$ is

$$\hat{\sigma} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{N}}$$

where N is the length of the text. Works better when

$$N\hat{\theta} \text{ and } N(1-\hat{\theta}) > 10$$

Approximate 95% confidence interval is

$$\hat{\theta} \;\pm\; 1.96\hat{\sigma}$$

Inference about proportions

Example: in the 2001 Labour manifesto there are 879 matches to Laver and Garry's +state category

- 0.029 (nearly 3%) of the document's words
- 0.093 (about 9%) of words that matched *any* categories

The document has 30825 words, so the *first* proportion is estimated as

$$\hat{\theta}_{+\text{state}} \;=\; 0.029 \;\; [0.027, 0.031]$$

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
  - That is probably not true, so it is probably *over*confident
- This is a quite general problem. . .

Reporting

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

Rates are more intuitive

The rate of dictionary matches per B words is

$$\lambda_B = \theta B$$

which is a more interpretable proportion.

Different measures correspond to different choices of B.

Reporting

Not all choices are constant or comparable across languages, documents and topics

Quantity	B	Constant?
Proportion	1	Yes
Word count	N	No
Block	B	Yes
Sentence	?	No
Paragraph	?	No

Under what circumstances are these measures comparable?

Inference about differences

The large sample standard error for $\hat{\theta}_i - \hat{\theta}_j$ is

$$\hat{\sigma} = \sqrt{\frac{\hat{\theta}_i(1-\hat{\theta}_i)}{N} + \frac{\hat{\theta}_j(1-\hat{\theta}_j)}{N}}$$

where N is the length of the text. Works better when

$$N\hat{\theta} \text{ and } N(1-\hat{\theta}) > 10$$

Approximate 95% confidence interval is

$$\hat{\theta}_i - \hat{\theta}_j \ \pm \ 1.96\hat{\sigma}$$

Inference about differences

UK Conservatives tend to target rural voters.

How much more attention did they get from the Conservatives than from Labour in 2001?

Consider the (very small) category 'rural'

Conservatives match 29 words, Labour 31, but Labour's manifesto is much longer so

$$\hat{\theta}^{\mathsf{LAB}} - \hat{\theta}^{\mathsf{CON}} \;=\; -0.0012 \;\; [-0.0003, -0.002]$$

This difference is significant (though see caveats above).

Inference about ratios

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

Compare instances of +state and -state in the manifestos

Party	Counts		Proportion	
	+S	-S	+S	-S
Conservative	386	880	.013	.03
Labour	439	390	.025	.022

Risk Ratios

Compute two *risk ratios*:

$$\begin{aligned} RR_{+\text{state}} &= \frac{P(+\text{state} \mid \text{cons})}{P(+\text{state} \mid \text{lab})} \\ RR_{\text{-state}} &= \frac{P(\text{-state} \mid \text{cons})}{P(\text{-state} \mid \text{lab})} \end{aligned}$$

and 95% confidence intervals

Risk Ratios

Standard error around estimated log *RR* is

$$\hat{\sigma} = \sqrt{\frac{1}{C_{\text{cons}}} - \frac{1}{N_{\text{cons}}} + \frac{1}{C_{\text{lab}}} - \frac{1}{N_{\text{lab}}}}$$

95% Confidence interval around log *RR* is

$$\log RR \ \pm \ 1.96\hat{\sigma}$$

Exponentiate the estimate and endpoints to get an interval for the risk ratio

Intepreting 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 language as the labour manifesto

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

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

Risk Ratios

	Risk Ratio
-S	1.35 [1.2, 1.5]
+S	0.53 [0.46,0.6]

Conservative manifesto generates 35% more -state words

- 35% = 100(1.35 - 1)%

Labour manifesto generates 89% more +state words

- 0.53 means *fewer* so
- 89% = 100(1/0.53 - 1)% more

Confidence interval suggests the increase is not less than 66% or more than 117%

Not doing it by hand

On the web, e.g.

- Use proportions, differences, equality tests: Vassar Stats
- For risk ratios: Calculator 3

Using the corpora library for R

```
> library(corpora)
> prop.cint(c, N) # interval(s) for proportion(s)
> chisq(c1, N1, c2, N2) # test for prop. equality
> rel.risk.cint(c1, N1, c2, N2) # a conservative RR
```

Handily, corpora functions tend to take vector arguments

What they probably didn't tell you

. . . in your statistics class

There is often more than one (reasonable) way to compute a confidence interval *particularly* with count data, e.g.

> *Newcombe R.G. (1998) Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. Statistics in Medicine 17, 857–872.*
>
> *Newcombe R.G. (1998) Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods. Statistics in Medicine 17, 873–890.*

Fortunately (or not) differences are usually smaller than *error due to our sampling assumptions*

More complex models

We have concentrated on reporting the immediate results of CCA, as proportions, rates, differences and ratios

Often you will want to use CCA output as a dependent variable in a larger analysis, e.g. a regression model

Options for when CCA output as *dependent* variable is essentially

- proportion data
- count data

Proportional data requires a (possibly overdispersed) Binomial assumption

Count data requires a (possibly overdispersed) Poisson assumption

Rate Example

Pro-independence language in Taiwanese presidential speeches (Sullivan and Lowe, forthcoming)

Chen was a

- Pro-independence *leader* (China, some defense analysts)
- Domestic audience *panderer* (DPP, some electoral analysts)

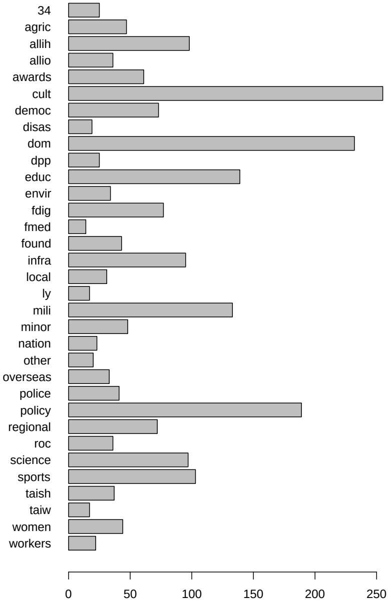
Previous work using in-depth qualitative analysis of ‘key speeches’ indicated strong pro-independence rhetoric

Rate Example

The two interpretations investigated using:

- Theory: Public opinion is flat, so rhetorical *increase* is due to leadership and rhetorical *variation* is due to audience
- Data: All presidential speeches ($\sim$ 3000)
- Measurement: Content analysis dictionary for pro-independence language
- Model: Rate of rhetoric as a function of time, events, with audience as a random effect

Audiences


34
agric
allih
allio
awards
cult
democ
disas
dom
dpp
educ
envir
fdig
fmed
found
infra
local
ly
mili
minor
nation
other
overseas
police
policy
regional
roc
science
sports
taish
taiw
women
workers
0
50
100
150
200
250

Rate Example

Model of independence category matches C per speech

$$\begin{aligned} C_{\text{indep}} &\sim \text{Poisson}(\lambda_{\text{indep}}) \\ E[\log\,\lambda_{\text{indep}}] &= \text{Const.} + \text{SecondTerm} + \text{HuWen} + \text{Law} + \log\left(\text{Total}\right) \end{aligned}$$

Const: the resting rate of pro-independence rhetoric

Second Term: dummy for presidential term

Hu and Wen: Dummy for new mainland leadership

Law: Anti-secession law passed

Total: Number of sentences in speech

Reminder

Poisson Distribution:

$$\begin{aligned} P(W_i) &= \text{Poisson}(\lambda_i) \\ &= \frac{\lambda_i^{W_i} \; e^{-\lambda}}{W_i!} \end{aligned}$$

Expected W_i is λ_i

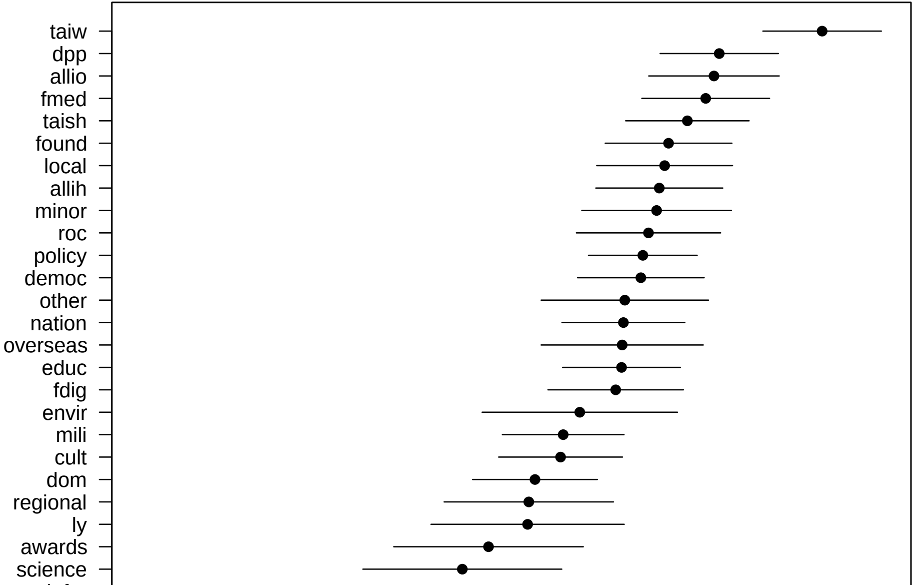
Variance of W_i is λ_i

Results: Fixed effects

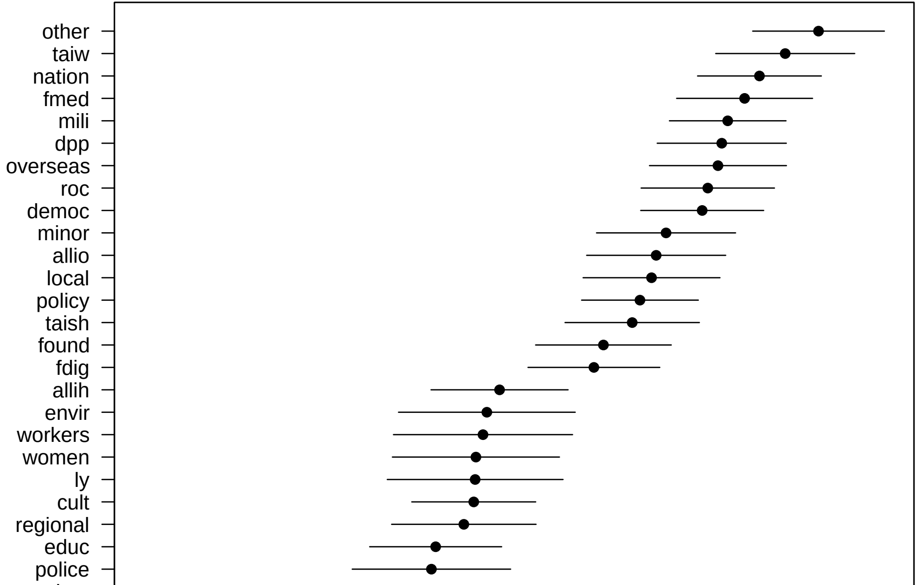
Fixed effects:

	Estimate	S.E.	z value	P(¿|z|)
Constant	-6.0137	0.2087	-28.8	¡.001
secondterm	0.1944	0.0558	3.5	¡.001
hw	0.1028	0.0444	2.3	¡.05
sec	0.0870	0.0483	1.8	.072

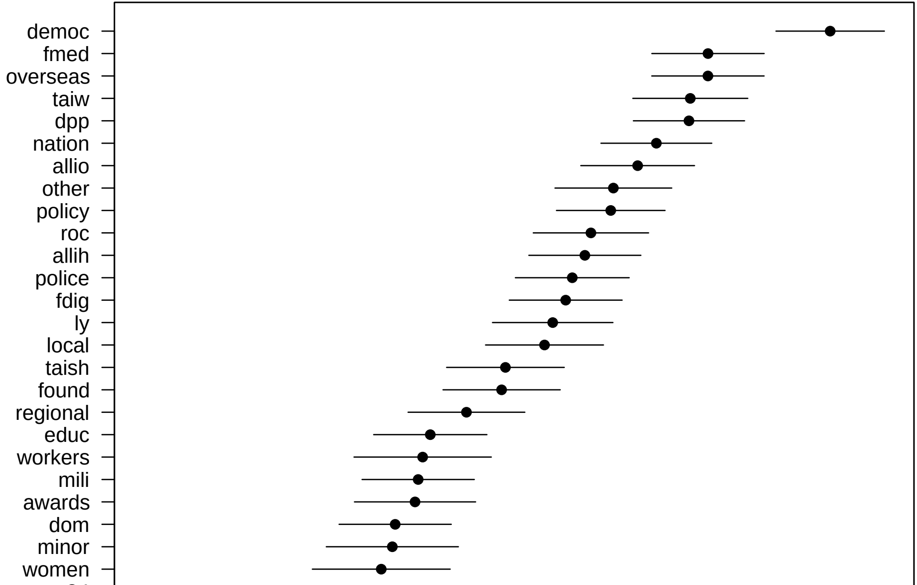
Results: Random effects



Other measures: Taiwan identity



Other measures: Pro-democracy



Background Theory

Some useful motivating theory (most in the Readings):

- Spatial theories (Laver and Garry, Riker)
- Salience and issue ownership (Budge et al.)
- Framing and reframing (e.g. Schoenhardt Bailey, Bara et al.)

Content Analysis Programs

Yoshikoder (Hamlet, Diction, Textpack, Wordstat, etc.)

LIWC (Linguistic Inquiry and Word Count, Pennebacker)

General Inquirer (Stone et al.)

Alceste (Image corp.)

See Lowe's review and also Alexa and Zuell (2000).

Content Analysis Programs

Yoshikoder is one of many classical content analysis programs having a basic handful of functions:

- Category building
- Concordance construction
- Frequency reports

Not as fancy as Wordstat but. . .

- free!
- works with non-english text
- works on all operating systems

Content Analysis Programs

LIWC is both a dictionary and a program (english only)

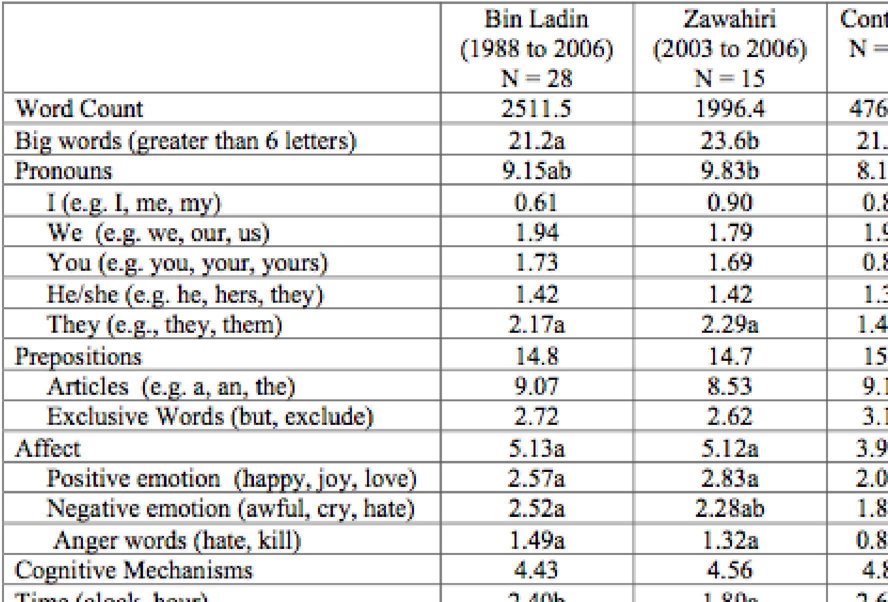
(one form of this dictionary is translated into Yoshikoder format and available from www.yoshikoder.org) Mostly used for social psychology applications

Has an online version

Example:

- Zawahiri vs. bin Laden vs. the world. . . (Pennebaker and Chung)

bin Laden vs. Zawahiri vs. Controls



	Bin Ladin (1988 to 2006) N = 28	Zawahiri (2003 to 2006) N = 15	Cont N =
Word Count	2511.5	1996.4	476
Big words (greater than 6 letters)	21.2a	23.6b	21.
Pronouns	9.15ab	9.83b	8.1
I (e.g. I, me, my)	0.61	0.90	0.8
We (e.g. we, our, us)	1.94	1.79	1.9
You (e.g. you, your, yours)	1.73	1.69	0.8
He/she (e.g. he, hers, they)	1.42	1.42	1.3
They (e.g., they, them)	2.17a	2.29a	1.4
Prepositions	14.8	14.7	15
Articles (e.g. a, an, the)	9.07	8.53	9.1
Exclusive Words (but, exclude)	2.72	2.62	3.1
Affect	5.13a	5.12a	3.9
Positive emotion (happy, joy, love)	2.57a	2.83a	2.0
Negative emotion (awful, cry, hate)	2.52a	2.28ab	1.8
Anger words (hate, kill)	1.49a	1.32a	0.8
Cognitive Mechanisms	4.43	4.56	4.8
Time (clock, hour)	2.40b	1.80a	2.6

Content Analysis Programs

The General Inquirer is perhaps the oldest content analysis program still in existence (1967)

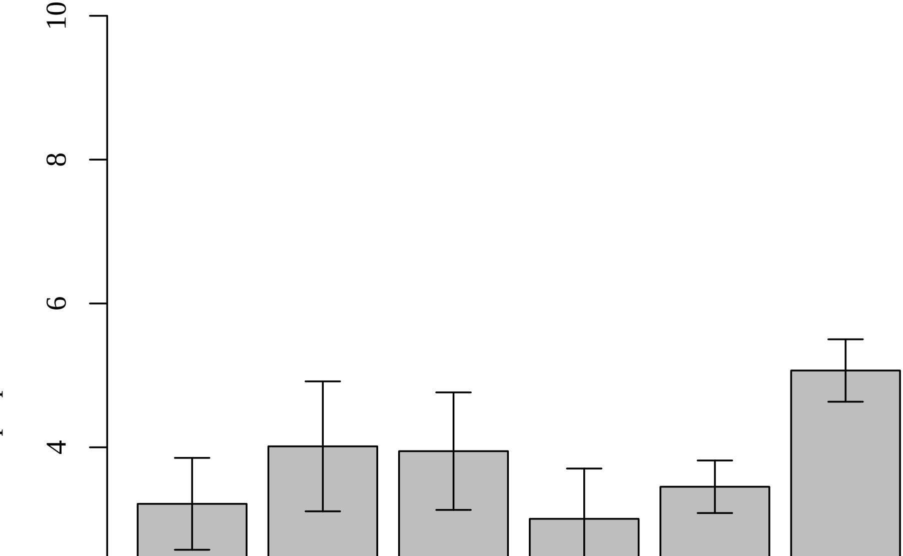
13000 words (and 6336 word sense disambiguation rules)

An online version is available at Maryland

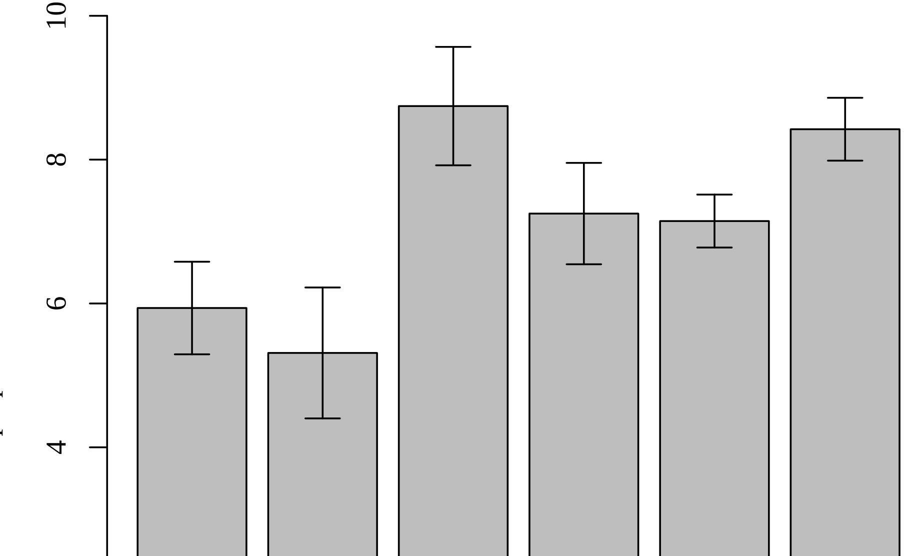
Example:

- speeches from US presidential candidates (2000)

Negative language



Positive language



How to build a dictionary

The ideal content analysis dictionary associates all and only the relevant words to each category in a perfectly valid scheme

Three issues:

Validity	Is the dictionary's category scheme valid?
Sensitivity	Does this dictionary identify *all* my content?
Specificity	Does it identify *only* my content?

How to build a dictionary

Assume you want to construct an entry for the category ‘Terrorism’

Imagine two different dictionary entries:

- One contains all the words in the language (D1)
- The other contains the word ‘terrorist’ (D2)

D1 is *highly sensitive*: no language about terrorism is ever missed, but *highly unspecific*: terrorism language is swamped

D2 is *highly specific*: the word occurs in discussions of terrorism, but *highly insensitive*: much terrorism language is ignored

Of course, useful dictionaries lie in the middle

How to build a dictionary

Different problems arise with more than one category, e.g.

- ‘Agricultural policy’ vs ‘National security’

Even if the categories *themselves* are exclusive there is always a chance a *word* suitable for one slips into the other category,

Or there are words that are used to describe both topics, e.g.

- ‘revolution’, ‘outbreak’, ‘quarantine’

That is a fact not easily dealt with by CCA. An explicitly statistical framework is needed.

A Sketch of the Statistical Framework

Assume $P(W \mid \theta)$ is

	θ	
	agriculture	security
nuclear	0	0.8
tractor	0.3	0
revolution	0.7	0.2
	1	1

A Sketch of the Statistical Framework

Bayes Theorem:

$$P(\theta \mid W) = \frac{P(W \mid \theta)P(\theta)}{P(W)}$$

So if $P(\theta = \text{`agriculture}') = 0.5$ then

θ

	agriculture	security	
nuclear	0	1	1
tractor	1	0	1
revolution	0.78	0.22	1

Proportions

Compute category proportions (as before):

$$\hat{\theta}_i = \frac{C_i}{\sum_j C_j}$$

C_i is a sum of $P(\theta = i \mid W)$s which can now be fractional

- e.g. two tokens of ‘revolution’ adds 1.56 to agriculture and 0.44 to security

How to build a dictionary

CCA requires that you deal with the specificity/sensitivity trade-off *yourself*

How to proceed?

Training, validation, and test sets

We can steal some useful terminology from Machine Learning:

Training set	documents you use to build the dictionary
Validation set	documents you use to tell how well you're doing
Test set	documents you use to quantify external validity

This scheme is intended to avoid 'over-fitting' — building a dictionary that is highly specific to a set of documents

A problem if you only sampled the population of texts, or want to use the dictionary on new data