Day 3: Dictionary Approaches

Kenneth Benoit

Quants 3: Quantitative Text Analysis

Week 3: March 9, 2018

Week 4 Outline

- Dictionary approach overview
- Some well-known dictionaries
- Advantages and disadvantages
- Dictionary construction
- Scaling dictionary results
- Keyword detection
- More complex models: beyond dictionaries

Bridging qualitative and quantitative text analysis

- A hybrid procedure between qualitative and quantitative classification the fully automated end of the text analysis spectrum
- "Qualitative" since it involves identification of the concepts and associated keys/categories, and the textual features associated with each key/category
- Dictionary construction involves a lot of contextual interpretation and qualitative judgment
- Perfect reliability because there is no human decision making as part of the text analysis procedure

"Dictionary": a misnomer?

- A *dictionary* is really a thesaurus: a canonical term or concept (a "key") associated with a list of equivalent synonyms
- But dictionaries tend to be exclusive: they single out features defined as keys, selecting the terms or patterns linked to each key
- An alternative is a "thesaurus" concept: a tag of key equivalency for an associated set of terms, but non-exclusive
  - WC = `wc, toilet, restroom, bathroom, jack, loo`
  - vote = `poll, suffrage, franchis*, ballot*, ^vot$`

Rationale for dictionaries

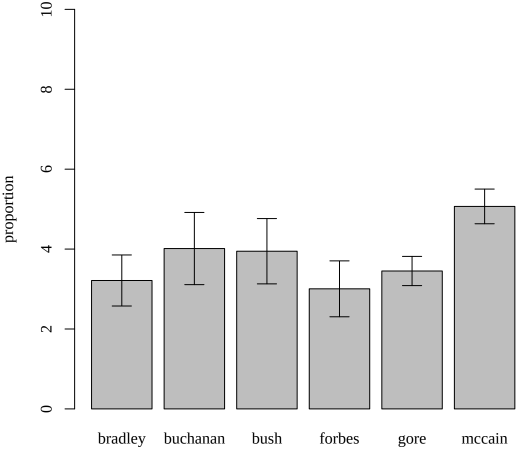
- Rather than count words that occur, pre-define words associated with specific meanings
- Two components:
  - **key** the label for the equivalence class for the concept or canonical term
  - **values** (multiple) terms or patterns that are declared equivalent occurences of the key class
- Frequently involves lemmatization: transformation of all inflected word forms to their "dictionary look-up form" — more powerful than stemming

Well-known dictionaries: General Inquirer

- General Inquirer (Stone et al 1966)
- Example: self = *I*, *me*, *my*, *mine*, *myself*
  selves = *we*, *us*, *our*, *ours*, *ourselves*
- Latest version contains 182 categories – the "Harvard IV-4" dictionary, the "Lasswell" dictionary, and five categories based on the social cognition work of Semin and Fiedler
- Examples: "self references", containing mostly pronouns; "negatives", the largest category with 2291 entries
- Also uses disambiguation, for example to distinguishes between *race* as a contest, *race* as moving rapidly, *race* as a group of people of common descent, and *race* in the idiom "rat race"
- Output example:
  `http://www.wjh.harvard.edu/~inquirer/Spreadsheet.html`

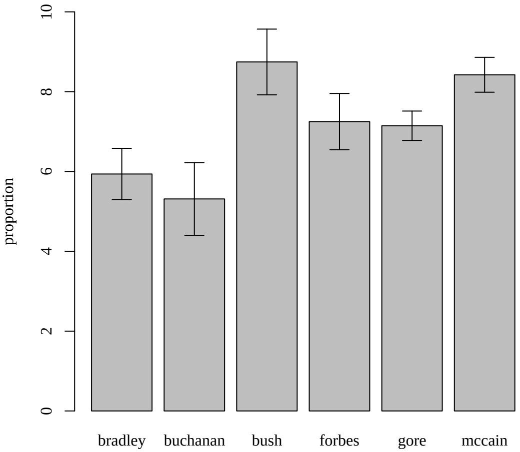
General Inquirer Applied to US Presidential Candidate Speeches (2000)

Negative language



General Inquirer Applied to US Presidential Candidate Speeches (2000)

Positive language



Well-known dictionaries: Regressive Imagery Dictionary

- Consists of about 3,200 words and roots, assigned to 29 categories of primary process cognition, 7 categories of secondary process cognition, and 7 categories of emotions
- designed to measure primordial vs. conceptual thinking
  - Conceptual thought is abstract, logical, reality oriented, and aimed at problem solving
  - Primordial thought is associative, concrete, and takes little account of reality – the type of thinking found in fantasy, reverie, and dreams
- Categories were derived from the theoretical and empirical literature on regressive thought by Martindale (1975, 1990)

Regressive Imagery Dictionary categories

- Full listing of categories

1 orality	21 brink-passage	41 aggression	62 novelty
2 anality	22 narcissism	42 expressive behaviour	63 negation
3 sex	23 concreteness	43 glory	64 triviality
4 touch	24 ascend	44 female role	65 transmute
5 taste	25 height	45 male fole	
6 odour	26 descent	46 self	
7 general sensation	27 depth	47 related others	
8 sound	28 fire	48 diabolic	
9 vision	29 water	49 aspiration	
10 cold	30 abstract thought	50 angelic	
11 hard	31 social behaviour	51 flowers	
12 soft	32 instrumental behaviour	52 synthesize	
13 passivity	33 restraint	53 streight	
14 voyage	34 order	54 weakness	
15 random movement	35 temporal references	55 good	
16 diffusion	36 moral imperative	56 bad	
17 chaos	37 positive affect	57 activity	
18 unknown	38 anxiety	58 being	
19 timelessness	39 sadness	59 analogy	
20 counscious	40 affection	61 integrative con	

- More on categories:
`http://www.kovcomp.co.uk/wordstat/RID.html`

Linquistic Inquiry and Word Count

- Created by Pennebaker et al — see `http://www.liwc.net`
- uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries
- For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. So observing the token *cried* causes each of these five subdictionary scale scores to be incremented
- Hierarchical: so "anger" are part of an *emotion* category and a *negative emotion* subcategory
- You can buy it here: `http://www.liwc.net/descriptiontable1.php`

Example: Terrorist speech

	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.

Example: Laver and Garry (2000)

- A *hierarchical* set of categories to distinguish policy domains and policy positions – similar in spirit to the CMP
- Five domains at the top level of hierarchy
  - economy
  - political system
  - social system
  - external relations
  - a " 'general' domain that has to do with the cut and thurst of specific party competition as well as uncodable pap and waffle"
- Looked for word occurences within "word strings with an average length of ten words"
- Built the dictionary on a set of specific UK manifestos

Example: Laver and Garry (2000): Economy

TABLE 1 Abridged Section of Revised Manifesto Coding Scheme

1 ECONOMY
Role of state in economy

- 1 1 ECONOMY/+State+
  Increase role of state
  - 1 1 1 ECONOMY/+State+/Budget
    Budget
    - 1 1 1 1 ECONOMY/+State+/Budget/Spending
      Increase public spending
      - 1 1 1 1 1 ECONOMY/+State+/Budget/Spending/Health
      - 1 1 1 1 2 ECONOMY/+State+/Budget/Spending/Educ. and training
      - 1 1 1 1 3 ECONOMY/+State+/Budget/Spending/Housing
      - 1 1 1 1 4 ECONOMY/+State+/Budget/Spending/Transport
      - 1 1 1 1 5 ECONOMY/+State+/Budget/Spending/Infrastructure
      - 1 1 1 1 6 ECONOMY/+State+/Budget/Spending/Welfare
      - 1 1 1 1 7 ECONOMY/+State+/Budget/Spending/Police
      - 1 1 1 1 8 ECONOMY/+State+/Budget/Spending/Defense
      - 1 1 1 1 9 ECONOMY/+State+/Budget/Spending/Culture
    - 1 1 1 2 ECONOMY/+State+/Budget/Taxes
      Increase taxes
      - 1 1 1 2 1 ECONOMY/+State+/Budget/Taxes/Income
      - 1 1 1 2 2 ECONOMY/+State+/Budget/Taxes/Payroll
      - 1 1 1 2 3 ECONOMY/+State+/Budget/Taxes/Company
      - 1 1 1 2 4 ECONOMY/+State+/Budget/Taxes/Sales
      - 1 1 1 2 5 ECONOMY/+State+/Budget/Taxes/Capital
      - 1 1 1 2 6 ECONOMY/+State+/Budget/Taxes/Capital gains
    - 1 1 1 3 ECONOMY/+State+/Budget/Deficit
      Increase budget deficit
      - 1 1 1 3 1 ECONOMY/+State+/Budget/Deficit/Borrow
      - 1 1 1 3 2 ECONOMY/+State+/Budget/Deficit/Inflation

Example: Laver and Garry (2000)

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

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

Advantage: Multi-lingual

APPENDIX B
DICTIONARY OF THE COMPUTER-BASED CONTENT ANALYSIS

	NL	UK	GE	IT
Core	elit*	elit*	elit*	elit*
	consensus*	consensus*	konsens*	consens*
	ondemocratisch*	undemocratic*	undemokratisch*	antidemocratic*
	ondemokratisch*			
	referend*	referend*	referend*	referend*
	corrupt*	corrupt*	korrupt*	corrot*
	propagand*	propagand*	propagand*	propagand*
	politici*	politici*	politiker*	politici*
	bedrog	*deceit*	täusch*	ingann*
	bedrieg	*deceiv*	betrüg*	
			betrug*	
	verraa	*betray*	*verrat*	tradi*
	verrad			
	schaam*	shame*	scham*	vergogn*
			schäm*	
	schand*	scandal*	skandal*	scandal*
	waarheid*	truth*	wahrheit*	verità
	oneerlijk*	dishonest*	unfair*	disonest*
			unehrlich*	
Context	establishm*	establishm*	establishm*	partitocrazia
	heersend*	ruling*	*herrsch*	
	capitul*			
	kapitul*			
	kaste*			
	leugen*		lüge*	menzogn*
	lieg*			mentir*

(from Rooduijn and Pauwels 2011)

Disdvantage: Highly specific to context

- Example: Loughran and McDonald used the Harvard-IV-4 TagNeg (H4N) file to classify sentiment for a corpus of 50,115 firm-year 10-K filings from 1994–2008
- found that almost three-fourths of the "negative" words of H4N were typically not negative in a financial context e.g. *mine* or *cancer*, or *tax*, *cost*, *capital*, *board*, *liability*, *foreign*, and *vice*
- Problem: polysemes – words that have multiple meanings
- Another problem: dictionary lacked important negative financial words, such as *felony*, *litigation*, *restated*, *misstatement*, and *unanticipated*

Different dictionary formats

- General Inquirer: see `http://www.wjh.harvard.edu/~inquirer/inqdict.txt`
- WordStat: see `http://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/`
- LIWC: for an example see the Moral Foundations dictionary at `http://www.moralfoundations.org/othermaterials`
- quanteda (see demo code)

A quick introduction to regular expressions

- an expanded version of the "glob" matching implemented in most command line interpreters, i.e.
  - `*` matches zero or more characters
  - `?` matches any one character (and in some environments, zero trailing characters)
  - `[]` may match any characters within a range inside the brackets
- a much more powerful version are *regular expressions*, which also exist in several (slightly) different versions
- R has both the POSIX 1003.2 and the Perl Compatible Regular Expressions implemented, see `?regex`
- Additional materials:
  - great cheat sheet
  - useful tutorial and reference

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 key issues:

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

- Imagine two logical extremes of including all words (too sensitive), or just one word (too specific)

Coding scheme fundamentals

1. First key principle: Hierarchy
   1.1 First level: Domain
   1.2 Second level: subdomain
   1.3 (Third+ levels: may be additional sub-domains)
2. Second key principle: Confrontation
   Lowest-level categories should be for/against pairs, or "for/neutral/against"
3. On testing: Not necessary at design stage in the same way as for human coding – this is replaced by sensitivity/specificity testing in dictionary construction

How to build a dictionary

1. Identify “extreme texts” with “known” positions. Examples:
   - Opposition leader and Prime Minister in a no-confidence debate
   - Opposition leader and Finance Minister in a budget debate
   - Five-star review of a product (excellent) and a one-star review (terrible)
2. Search for differentially occuring words using word frequencies
3. Examine these words in context to check their sensitivity and specificity
4. Examine inflected forms to see whether stemming or wildcarding is required
5. Use these words (or their lemmas) for categories

Detecting "keywords"

- Detects words that *discriminate* between partitions of a corpus
- For instance, we could partition the Irish budget speech corpus into "government" and "opposition" speeches, and look for words that occur in one partition with higher relative frequency in opposition than in government speeches
- This is done by constructing a 2×2 table for each word, and testing association between that word and the partition categories

Detecting "keywords": Constructing the association table

	Target	~ Target	
Word 1	n_{11}	n_{12}	$n_{1.}$
~ (Word 1)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

- Once this is constructed, any standard measures of association (similar to those used to detect collocations) can be used to identify keyword associations with a class
- Same association measures are used as with collocation detection

statistical association measures

where m_{ij} represents the cell frequency expected according to independence:

G^2 likelihood ratio statistic, computed as:

$$2 * \sum_i \sum_j (n_{ij} * log \frac{n_{ij}}{m_{ij}}) \tag{1}$$

χ^2 Pearson's χ^2 statistic, computed as:

$$\sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{2}$$

statistical association measures (cont.)

pmi point-wise mutual information score, computed as $\log n_{11}/m_{11}$

dice the Dice coefficient, computed as

$$\frac{n_{11}}{n_{1.} + n_{.1}} \tag{3}$$

Examples

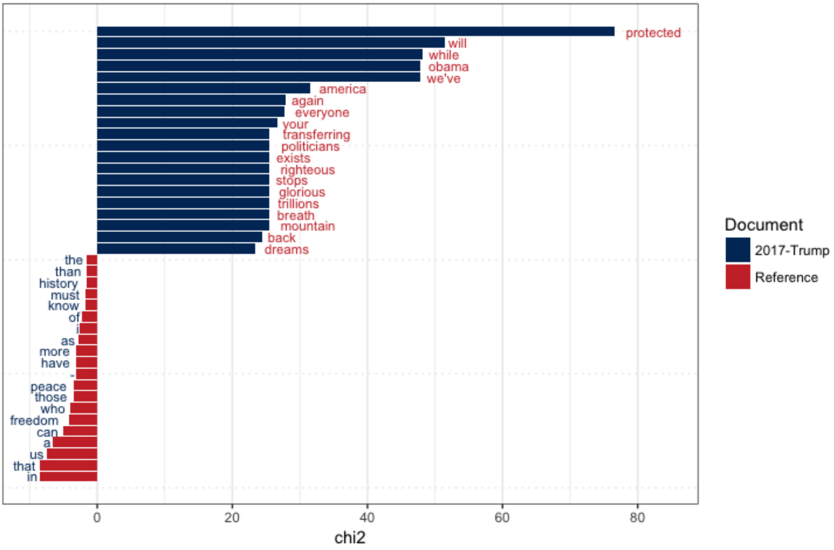
```
# compare Trump 2017 to other post-war preseidents
period <- ifelse(docvars(data_corpus_inaugural, "Year") < 1945,
                 "pre-war", "post-war")
pwdfm <- dfm(corpus_subset(data_corpus_inaugural, period == "post-war"))

textstat_keyness(pwdfm, target = "2017-Trump") %>%
    head(n = 7)
#     feature     chi2            p n_target n_reference
# 1 protected 76.64466 0.000000e+00        5           1
# 2      will 51.44795 7.351897e-13       40         299
# 3     while 48.23022 3.790079e-12        6           7
# 4     obama 47.85727 4.584000e-12        3           0
# 5     we've 47.85727 4.584000e-12        3           0
# 6   america 31.45537 2.040775e-08       18         112
# 7     again 27.81145 1.337322e-07        9          33
```

Examples

```
# using the likelihood ratio method
textstat_keyness(dfm_smooth(pwdfm), measure = "lr", target = "2017-Trum
    head()
#    feature        G2            p n_target n_reference
# 1     will 24.604106 7.040156e-07       41         317
# 2  america 14.040255 1.789387e-04       19         130
# 3     your 10.435140 1.236402e-03       12          68
# 4    again  9.758516 1.784939e-03       10          51
# 5    while  9.504990 2.049139e-03        7          25
# 6 american  8.877690 2.886766e-03       12          76


textstat_keyness(pwdfm, target = "2017-Trump") %>%
    textplot_keyness()
```



protected
will
while
obama
we've
america
again
everyone
your
transferring
politicians
exists
righteous
stops
glorious
trillions
breath
mountain
back
dreams
the
than
history
must
know
of
i
as
more
have
-
peace
those
who
freedom
can
a
us
that
in
0
20
40
60
80
chi2
Document
2017-Trump
Reference

Examples

Table 5
Keywords by gender in interview text: Selected categories[a]

Prostate	Breast
Treatment	
Catheter, brachytherapy, hormone, Zoladex, treatment, seeds, prostatectomy, Casodex, injection, radiation, injections, operation, Viagra, beam, radical, bag, Spes, Flutamide, tubes, capsule, Prazosin, tablets, watchful [waiting], cryosurgery, cryotherapy, Muse, probes, [watchful] waiting, therapy, strapped	Chemotherapy, Tamoxifen, mastectomy, prosthesis, chemo, lumpectomy, needle, HRT, scar, drains
Support	
NO KEYWORDS	Help, supportive, support, helped
Feelings	
Concerned, embarrassment	Feel, felt, want, need, cope, scared, crying, ups [and downs], wanted, depressed, scary, brave, cried, angry, coping, coped, feelings, fight, hard, upset
People	
Wife, he, men, man, chap, male, his, chaps, guy	I, she, husband, her, you, women, my, people, mum, sister, everybody, me, children, mother, friends, woman, lady, dad, she'd, daughter, she's, yourself, myself, sisters, I'd, auntie, ladies, who've, someone, somebody, your
Superlatives	
NO KEYWORDS	Wonderful, lovely, lots, amazing, marvellous

[a]Each section lists words in descending order of 'keyness'; 'split' words are excluded.

What to do with dictionary results

- Describe the results
- Scale quantities: pro- v. anti-, left v. right, etc. Example: Laver and Garry (see Lowe et al 2011 for alternatives)
- Could use these as features to measure similarity using (e.g.) cosine similarity
- Treat as other features and use machine learning or data mining methods

Scaling Issues

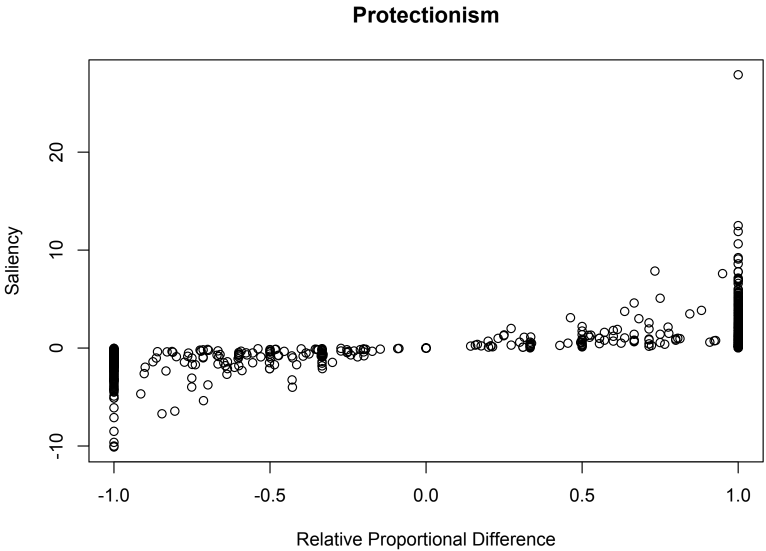
- Scaling becomes a major issue when we wish to construct quantities of interest from quantitative content analyses
- Simple example: Proportion of content of a given type (e.g. anti-Lisbon treaty)
- Complex example: Left-right policy positions (e.g. CMP "Rile")
- Are the metrics "natural"?
- Does the output metric resemble the input metric (if any)?
- What properties should the scale have, such as boundaries, type of increase, etc?
- How can uncertainty be characterized for the given scale?

Logit scale for left-right

- The Comparative Manifesto Project scales policy positions as absolute porportional difference, measured by proportion of "Right" mentions less proportion of "Left" mentions: $\frac{(R-L)}{N}$
- Problems:
  - Addition of irrelevant content shifts the scale toward zero
  - Assumes the additional mentions increase emphasis in a linear scale
- The alternative is to scale $\frac{(R-L)}{(R+L)}$ (Kim and Fording 2002; Laver and Garry 2000), but this too has problems:
  - Still linear shift in position for increase in repetition
  - Quickly maxes out at the extremes
- Lowe, Benoit, Mikhaylov and Laver (2010) propose using a logistic odds-ratio scale $\log \frac{R}{L}$

Comparing scales:

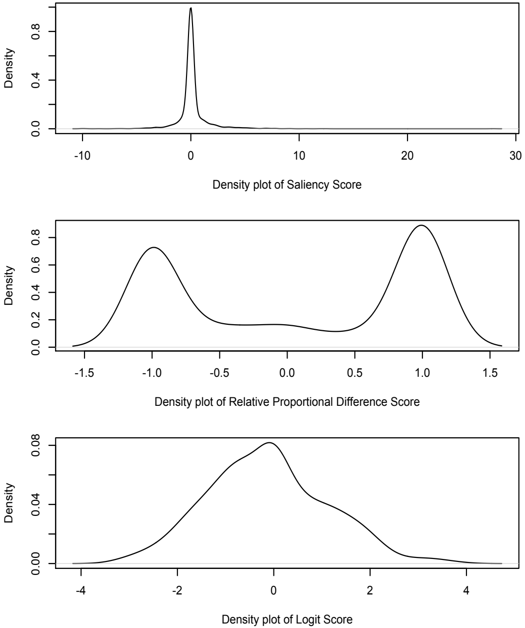
$\hat{\theta}^{(S)}$ v. $\hat{\theta}^{(R)}$



Comparing scales

Protectionism

distributions



More complex models

- More complex models are possible, when word rate occurrence is modeled more directly
- Example: Word rate occurrence could be Poisson distributed, and the dictionary approach simply selects specific words by pre-identified features
- From the quantitative matrix of (for instance) dictionary word occurrences by document, it would be possible to apply more advanced scaling or measurement methods
- But our next generalization will not involve modelling word rates by focusing on their stochastic process, but rather focusing on a relative probability model of word occurrence given a specific orientation

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