Day 6: Automated dictionary-based approaches

Kenneth Benoit

Essex Summer School 2011

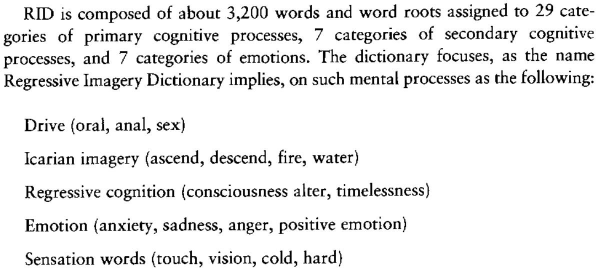
July 15, 2011

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:

	$P(\theta =$ 'Pro-State$'$ $\mid$ W$)$	$P(\theta =$ 'Anti-State$'$ $\mid$ 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. . .

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...

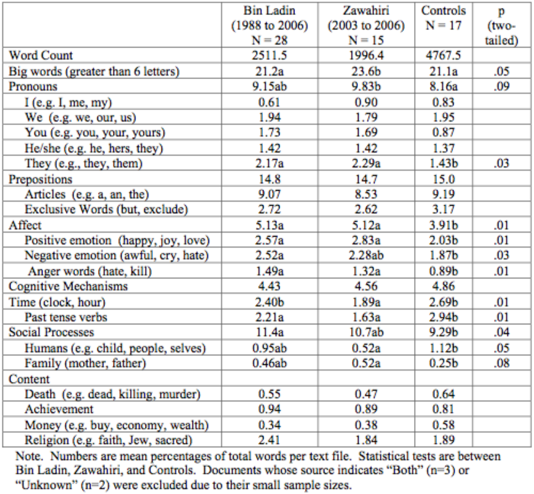
- free!
- works with non-english text
- works on all operating systems

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	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	

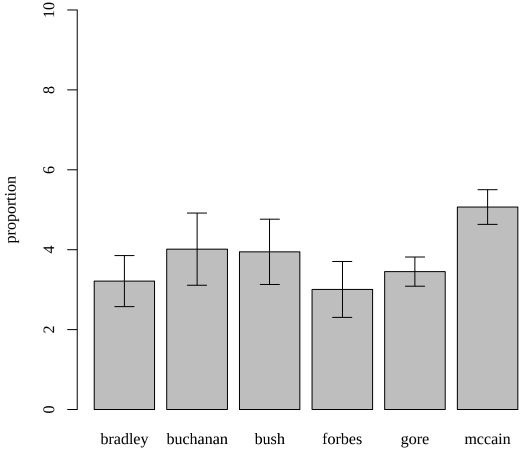
Note. Numbers are mean percentages of total words per text file. Statistical tests are between Bin Ladin, Zawahiri, and Controls. Documents whose source indicates "Both" (n=3) or "Unknown" (n=2) were excluded due to their small sample sizes.

Content Analysis Programs

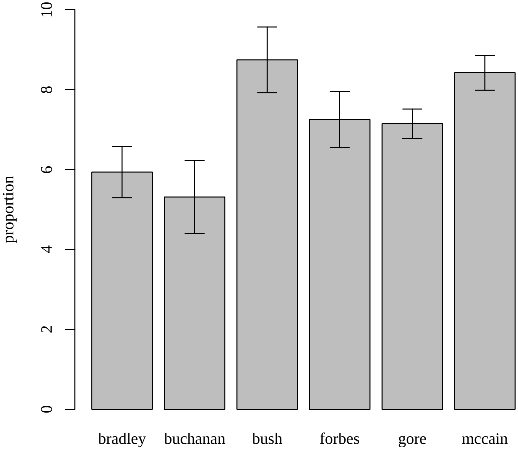
The General Inquirer is perhaps the oldest content analysis program still in existence (1967)
13000 words (and 6336 word sense disambiguation rules)
An online version is available at Maryland
Example:

- 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.