Day 6: Working with Textual Data

Kenneth Benoit

Data Mining and Statistical Learning

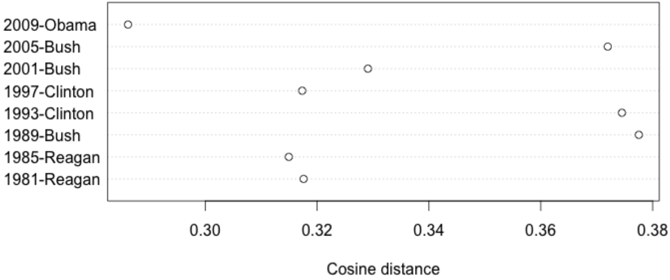
March 23, 2015

Distance measures

```r
library(proxy, warn.conflicts = FALSE, quietly = TRUE)
summary(pr_DB)

## * Similarity measures:
## Braun-Blanquet, Chi-squared, correlation, cosine, Cramer, Dice,
## eJaccard, Fager, Faith, Gower, Hamman, Jaccard, Kulczynski1,
## Kulczynski2, Michael, Mountford, Mozley, Ochiai, Pearson, Phi,
## Phi-squared, Russel, simple matching, Simpson, Stiles, Tanimoto,
## Tschuprow, Yule, Yule2
##
## * Distance measures:
## Bhjattacharyya, Bray, Canberra, Chord, divergence, Euclidean,
## fJaccard, Geodesic, Hellinger, Kullback, Levenshtein, Mahalanobis,
## Manhattan, Minkowski, Podani, Soergel, supremum, Wave, Whittaker
```

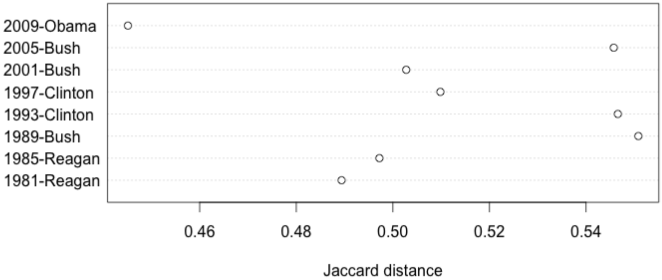
Example: Inaugural speeches, cosine distance to Obama 2014

```
library(quanteda)
presDfm <- dfm(subset(inaugCorpus, Year>1980),
               ignoredFeatures=stopwords("english", verbose=FALSE),
               stem=TRUE, verbose=FALSE)
obamaDistance <- as.matrix(dist(as.matrix(presDfm), "Cosine"))
dotchart(obamaDistance[1:8,9], xlab="Cosine distance")
```



Example: Jaccard distance to Obama

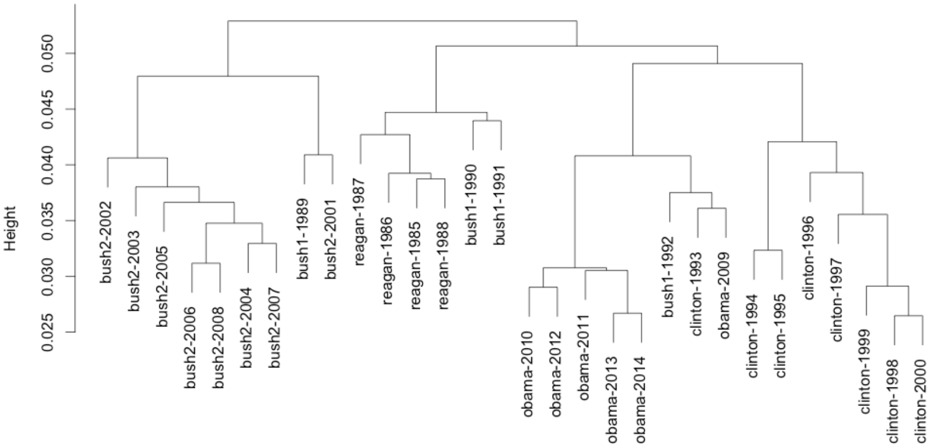
```
obamaDistance <- as.matrix(dist(as.matrix(presDfm), "eJaccard"))
dotchart(obamaDistance[1:8,9], xlab="Jaccard distance")
```



Dendrogram: Presidential State of the Union addresses

```
data(SOTUCorpus, package="quantedaData")
presDfm <- dfm(subset(SOTUCorpus, year>1960), verbose=FALSE, stem=TRUE,
               ignoredFeatures=stopwords("english", verbose=FALSE))
presDfm <- trim(presDfm, minCount=5, minDoc=3)

## Features occurring less than 5 times: 4079
## Features occurring in fewer than 3 documents: 3524

# hierarchical clustering - get distances on normalized dfm
presDistMat <- dist(as.matrix(weight(presDfm, "relFreq")))
# hiarchical clustering the distance object
presCluster <- hclust(presDistMat)
# label with document names
presCluster$labels <- docnames(presDfm)
# plot as a dendrogram
plot(presCluster)
```

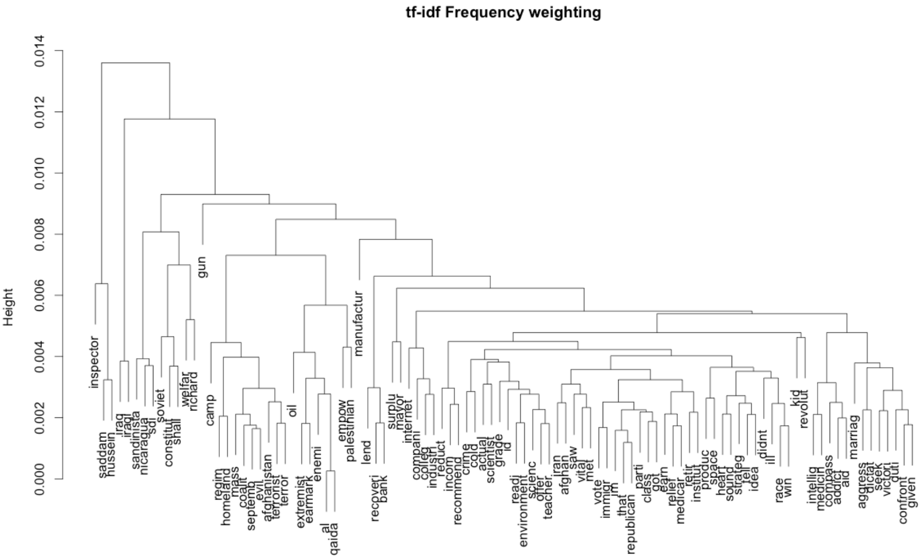
Dendrogram: Presidential State of the Union addresses



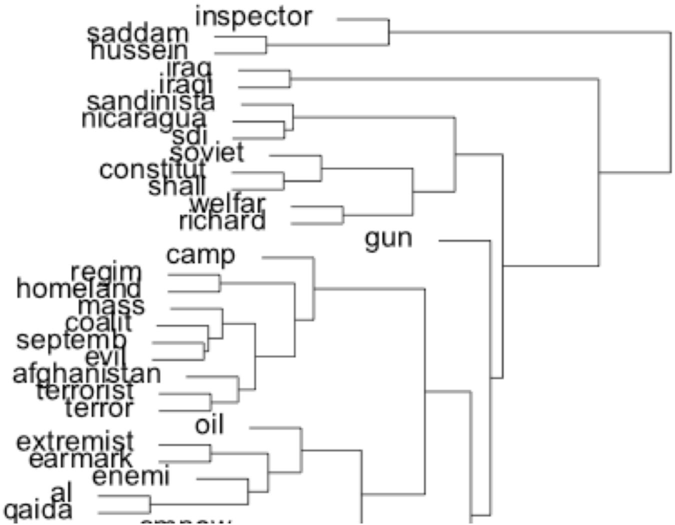
Dendrogram: Presidential State of the Union addresses

```
# word dendrogram with tf-idf weighting
wordDfm <- sort(tfidf(presDfm))  # sort in decreasing order of total word freq
wordDfm <- t(wordDfm)[1:100,]  # because transposed
wordDistMat <- dist(wordDfm)
wordCluster <- hclust(wordDistMat)
plot(wordCluster, xlab="", main="tf-idf Frequency weighting")
```

Dendrogram: Presidential State of the Union addresses



Dendrogram: Presidential State of the Union addresses



Singular Value Decomposition

- A matrix $\underset{i \times j}{\mathbf{X}}$ can be represented in a dimensionality equal to its rank k as:

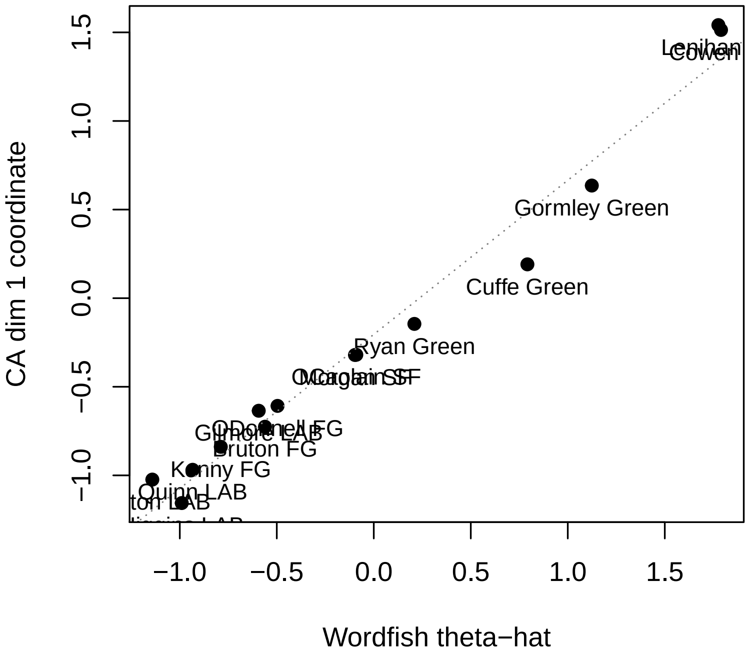
$$\underset{i \times j}{\mathbf{X}} = \underset{i \times k}{\mathbf{U}} \; \underset{k \times k}{\mathbf{d}} \; \underset{j \times k}{\mathbf{V}'} \tag{1}$$

- The **U**, **d**, and **V** matrixes "relocate" the elements of **X** onto new coordinate vectors in n-dimensional Euclidean space
- Row variables of **X** become points on the **U** column coordinates, and the column variables of **X** become points on the **V** column coordinates
- The coordinate vectors are perpendicular (*orthogonal*) to each other and are normalized to unit length

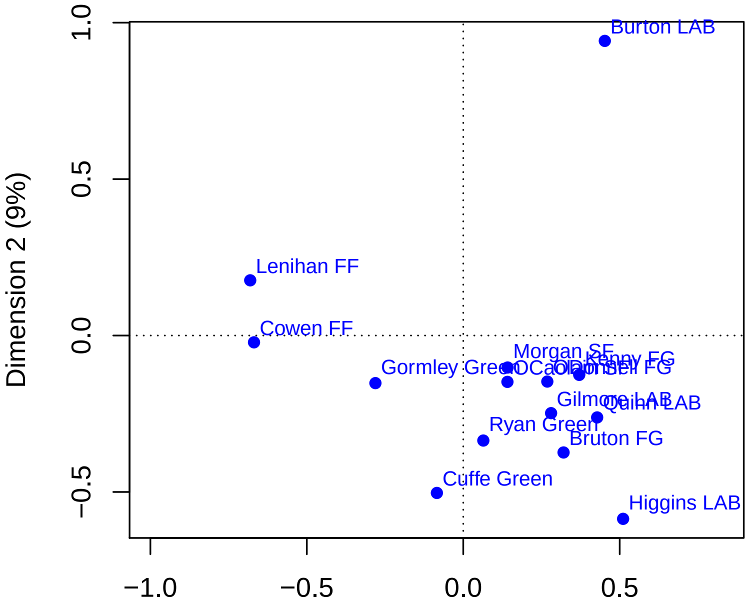
Correspondence Analysis and SVD

- Divide each value of $\mathbf{X}$ by the geometric mean of the corresponding marginal totals (square root of the product of row and column totals for each cell)
  - Conceptually similar to subtracting out the χ^2 expected cell values from the observed cell values
- Perform an SVD on this transformed matrix
  - This yields singular values $\mathbf{d}$ (with first always 1.0)
- Rescale the row ($\mathbf{U}$) and column ($\mathbf{V}$) vectors to obtain canonical scores (rescaled as $U_i\sqrt{f_{..}/f_{i.}}$ and $V_j\sqrt{f_{..}/f_{j.}}$)

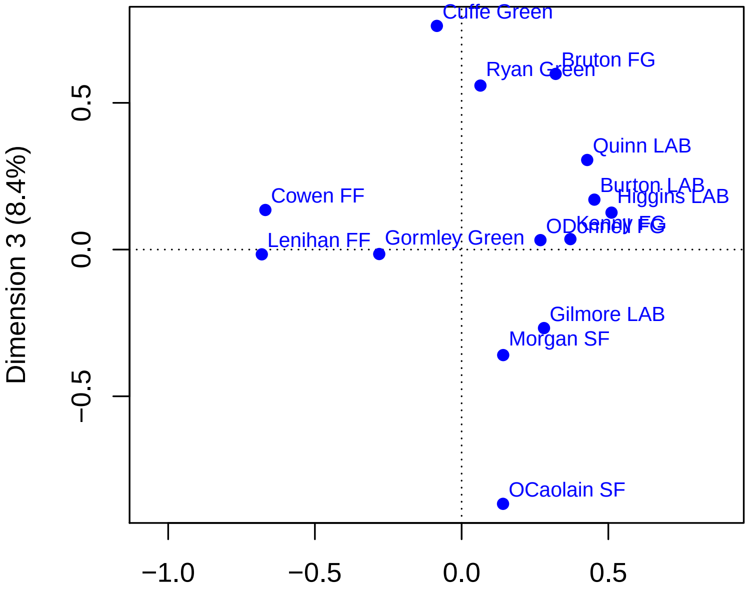
```r
data(ie2010Corpus, package="quantedaData")
# make prettier document names
docnames(ie2010Corpus) <-
    paste(docvars(ie2010Corpus, "name"), docvars(ie2010Corpus, "party"))
ieDfm <- dfm(ie2010Corpus)

## Creating a dfm from a corpus ...
##    ... indexing 14 documents
##    ... tokenizing texts, found 49,738 total tokens
##    ... cleaning the tokens, 845 removed entirely
##    ... summing tokens by document
##    ... indexing 4,859 feature types
##    ... building sparse matrix
##    ... created a 14 x 4859 sparse dfm
##    ... complete. Elapsed time: 0.712 seconds.

wf <- textmodel_wordfish(ieDfm, dir=c(2,1))
wca <- textmodel_ca(ieDfm)
plot(wf@theta, -1*wca$rowcoord[,1],
     xlab="Wordfish theta-hat", ylab="CA dim 1 coordinate", pch=19)
text(wf@theta, -1*wca$rowcoord[,1], docnames(ieDfm), cex=.8, pos=1)
abline(lm(-1*wca$rowcoord[,1] ~ wf@theta), col="grey50", lty="dotted")
```


Lenihan
Cowen
Gormley Green
Cuffe Green
Ryan Green
OCaolain SF
Morgan SF
ODonnell FG
Gilmore LAB
Bruton FG
Kenny FG
Quinn LAB
1.5
1.0
0.5
0.0
−0.5
−1.0
CA dim 1 coordinate
−1.0
−0.5
0.0
0.5
1.0
1.5
Wordfish theta−hat

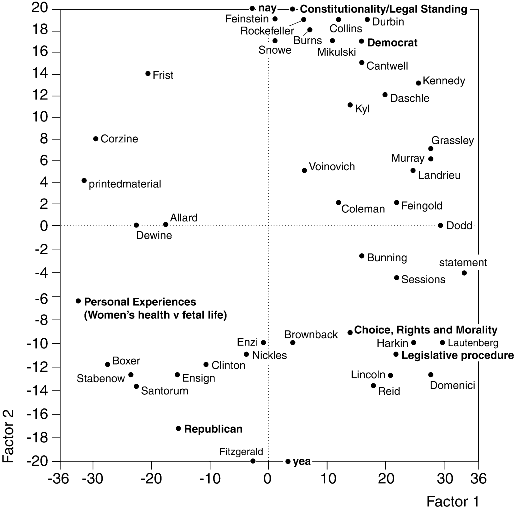
Dimension 1 v. Dimension 2



Dimension 1 v. Dimension 3



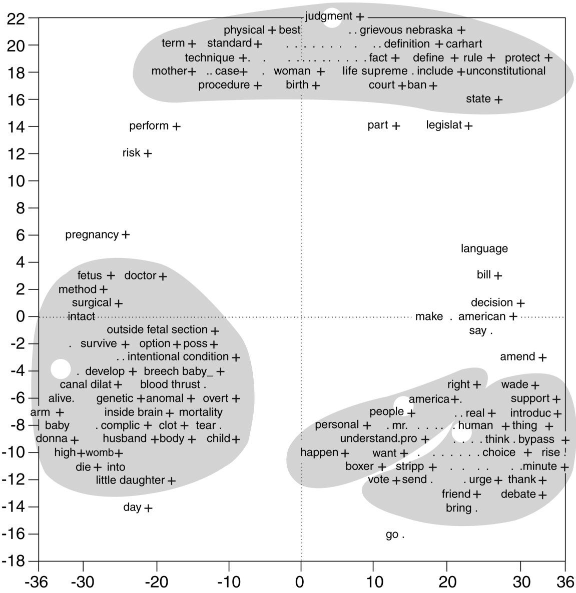
Example: Schonhardt-Bailey (2008) - speakers



	Eigenvalue	% Association	% Cumulative
Factor 1	0.30	44.4	44.4
Factor 2	0.22	32.9	77.3

Fig. 3. Correspondence analysis of classes and tags from Senate debates on Partial-Birth Abortion Ban Act

Example: Schonhardt-Bailey (2008) - words



The Poisson scaling "wordfish" model

Data:

- Y is N (speaker) $\times$ V (word) term document matrix
  $V \gg N$

Model:

$$
\begin{aligned}
P(Y_i \mid \theta) &= \prod_{j=1}^{V} P(Y_{ij} \mid \theta_i) \\
Y_{ij} &\sim \text{Poisson}(\lambda_{ij}) \qquad \text{(POIS)} \\
\log \lambda_{ij} &= (\log)\, \alpha_i + \theta_i \beta_j + \psi_j
\end{aligned}
$$

Estimation:

- Easy to fit for large V (V Poisson regressions with α offsets)

Model components and notation

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

Element	*Meaning*
i	indexes documents
j	indexes word types
θ_i	the unobservable "position" of document i
β_j	word parameters on θ – the relationship of word j to document position
ψ_j	word "fixed effect" (function of the frequency of word j)
α_i	document "fixed effects" (a function of (log) document length to allow estimation in Poisson of an essentially multinomial process)

How to account for uncertainty

- Ignore the problem and hope it will go away
  - SVD-based methods (e.g. correspondence analysis) typically do not present errors
  - and traditionally, point estimates based on other methods have not either
- Analytical derivatives
  - The covariance matrix is (asymptotically) the inverse of the negative of the Hessian
    (where the negative Hessian is the observed Fisher information matrix, a.ka. the second derivative of the log-likelihood evaluated at the maximum likelihood estimates)
  - Problem: These are *too small*
- Posterior sampling from MCMC

How to account for uncertainty (cont.)

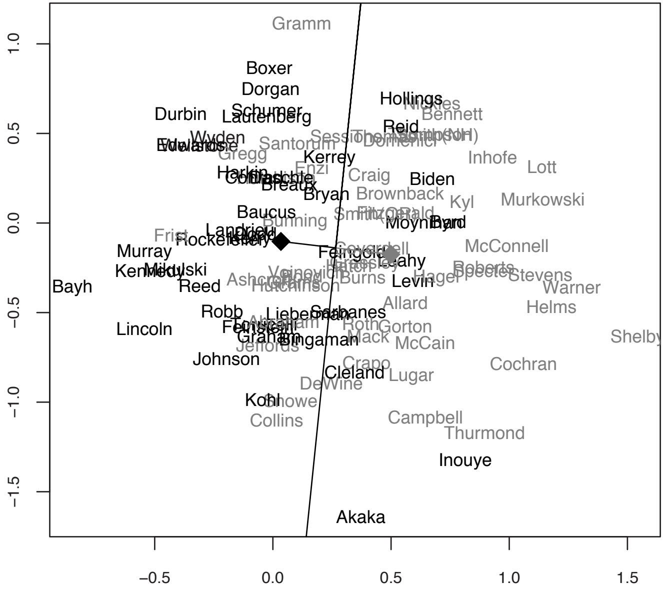
- Parametric bootstrapping (Slapin and Proksch, Lewis and Poole)
  Assume the distribution of the parameters, and generate data after drawing new parameters from these distributions.
- Non-parametric bootstrapping
  - draw new versions of the texts, refit the model, save the parameters, average over the parameters

Dimensions

How infer more than one dimension?
This is two questions:

- How to get two dimensions (for all policy areas) at the same time?
- How to get one dimension for each policy area?

The hazards of ex-post interpretation illustrated



Interpreting scaled dimensions

- In practice can be very subjective, involves interpretation
- Another (better) option: compare them other known descriptive variables
- Hopefully also *validate* the scale results with some human judgments
- This is necessary even for single-dimensional scaling
- And just as applicable for non-parametric methods (e.g. correspondence analysis) as for the Poisson scaling model

Using 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

"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$`

Bridging qualitative and quantitative text analysis

- A hybrid procedure between qualitative and quantitative classification the fully automated end of the text analysis spectrum
- “Qualitiative” 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

Linquistic Inquiry and Word Count

- Craeted 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