Computer-assisted content analysis

Will Lowe Princeton University
James Lo University of Southern California

Menu

Session 0: How could this possibly work?

Session 1: Dictionary-based 'classical' content analysis and topic models

Session 2: Classification and evaluation

Session 3: Scaling models

- Documents in space
- Modeling relative emphasis
- Validating human judgement
- Dimensionality

Documents in (ideological) space

	neue	vor	Menschen	wie	nur	Arbeitsplätze	…
...							
FDP-2005	11	20	6	22	31	17	…
FDP-2002	17	17	27	30	35	9	…
PDS-2005	5	10	17	10	9	12	…
PDS-2002	15	19	8	9	3	9	…
GREENS-2005	42	21	47	46	19	17	…
GREENS-2002	27	18	27	28	22	21	…
SPD-2005	8	15	26	11	13	10	…
SPD-2002	16	18	16	16	9	7	…
CDU-2005	21	12	10	13	19	22	…
CDU-2002	20	20	14	15	18	7	…
...							

Manifestos as bags of words

Documents in (ideological) space

e.g. the CMP (Budge et al. 1983).

Topic code	Meaning
403	Market Regulation
404	Economic Planning
405	Corporatism
⋮	
601	National Way of Life: Positive
602	National Way of Life: Negative
603	Traditional Morality: Positive
604	Traditional Morality: Negative
605	Law and Order

Documents in (ideological) space

e.g. the CMP (Budge et al. 1983).

	201	202	403	404	405	601	…
...							
FDP-1990	2	19	28	0	0	0	…
FDP-1994	0	11	17	0	0	0	…
FDP-1998	6	0	8	0	10	20	…
FDP-2002	26	11	31	1	0	10	…
FDP-2005	12	27	55	8	0	7	…
FDP-2009	10	38	16	21	0	10	…
...							

Manifestos as bags of topics

Back to the (contingency) table

Recall our two assumptions:

- A matrix of document by word/topic counts is a *contingency table* generated by unobserved *positions*
- Words occur at a *rate* determined by the content they are being used to express

$$C_{ij} \sim \text{Poisson}(\lambda_{ij})$$

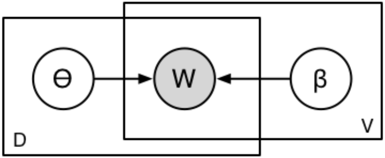
Back to the (contingency) table

How to connect the rates of each word in a document to θs (and βs)

	neue	vor	Menschen	wie	nur	Arbeitsplätze	…	
...								
FDP-2005	11	20	6	22	31	17	…	$\theta_{\text{FDP-2005}}$
FDP-2002	17	17	27	30	35	9	…	$\theta_{\text{FDP-2002}}$
PDS-2005	5	10	17	10	9	12	…	$\theta_{\text{PDS-2005}}$
PDS-2002	15	19	8	9	3	9	…	$\theta_{\text{PDS-2002}}$
GREENS-2005	42	21	47	46	19	17	…	$\theta_{\text{GREENS-2005}}$
GREENS-2002	27	18	27	28	22	21	…	$\theta_{\text{GREENS-2002}}$
SPD-2005	8	15	26	11	13	10	…	$\theta_{\text{SPD-2005}}$
SPD-2002	16	18	16	16	9	7	…	$\theta_{\text{SPD-2002}}$
CDU-2005	21	12	10	13	19	22	…	$\theta_{\text{CDU-2005}}$
CDU-2002	20	20	14	15	18	7	…	$\theta_{\text{CDU-2002}}$
...								
	β_{neue}	β_{vor}	β_{Menschen}	β_{wie}	β_{nur}	$\beta_{\text{Arbeitsplätze}}$		

Back to the (contingency) table

How to connect the rates of each word in a document to θs (and βs)



Simple models of count data

There are two *log-linear models* of any contingency table

$$\begin{aligned} \log \mu_{ij} &= \alpha_i + \psi_j && \text{(boring)} \\ &= \alpha_i + \psi_j + \lambda_{ij} && \text{(pointless)} \end{aligned}$$

Where's the relative emphasis?

Two models: There are two *log-linear models* of any contingency table

$$\begin{aligned} \log \mu_{ij} &= \alpha_i + \psi_j && \text{(independence)} \\ &= \alpha_i + \psi_j + \lambda_{ij} && \text{(saturated)} \end{aligned}$$

All the *relative emphasis* (and all the *political position-taking*) is in λ

Scaling models give dimensional structure to λ.

Infer dimensional structure

Intuition: λ has an orthogonal decomposition

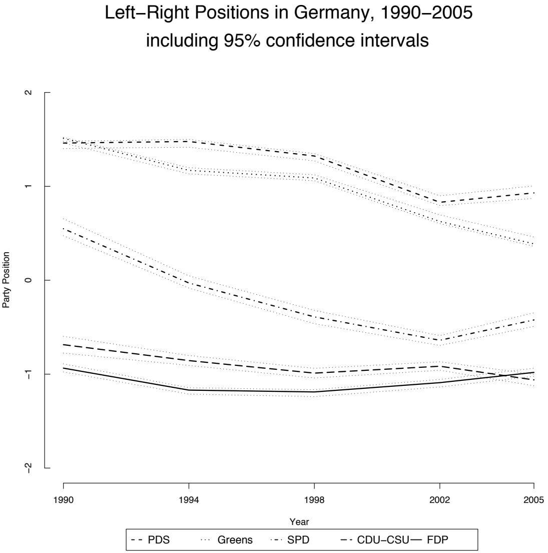
$$\begin{aligned}
\lambda &= \Theta \Sigma B^T && \text{(SVD)} \\
&= \sum_{m}^{M} \theta_{(m)} \sigma_{(m)} \beta_{(m)}^T \\
&\approx \theta \, \sigma \, \beta^T && \text{(Rank 1 approx.)}
\end{aligned}$$

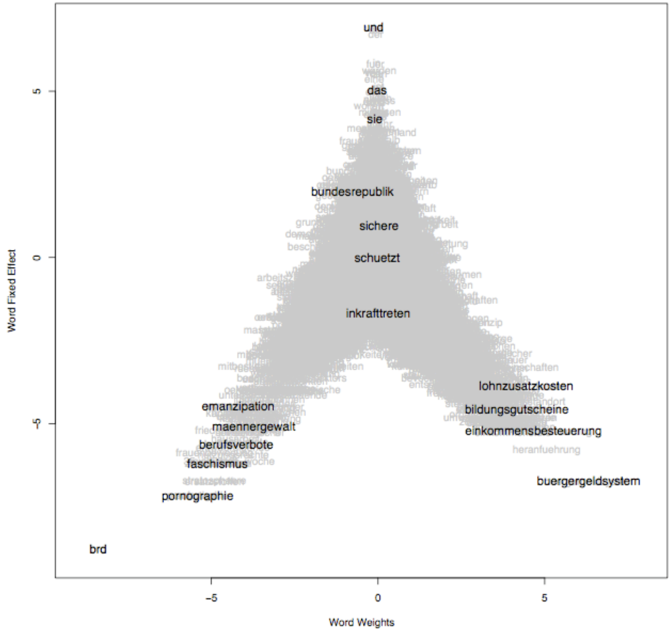
Infer dimensional structure

Intuition: λ has an orthogonal decomposition

$$\begin{aligned} \lambda &= \Theta \Sigma B^T && \text{(SVD)} \\ &= \sum_m^M \theta_{(m)} \sigma_{(m)} \beta_{(m)}^T \\ &\approx \theta\, \sigma\, \beta^T && \text{(Rank 1 approx.)} \end{aligned}$$

θ are *document positions*, β are *word positions*


Left−Right Positions in Germany, 1990−2005
including 95% confidence intervals
Party Position
2
1
0
−1
−2
1990
1994
1998
2002
2005
Year
PDS
Greens
SPD
CDU−CSU
FDP



und
der
fuer
werden
das
sie
bundesrepublik
sichere
schuetzt
inkrafttreten
lohnzusatzkosten
bildungsgutscheine
einkommensbesteuerung
heranfuehrung
buergergeldsystem
emanzipation
maennergewalt
berufsverbote
faschismus
pornographie
brd
Word Fixed Effect
5
0
-5
Word Weights
-5
0
5

Infer dimensional structure

Intuition: λ has a orthogonal decomposition

$$
\begin{aligned}
\lambda &= \Theta \Sigma B^T && \text{(SVD)} \\
&= \sum_{m}^{M} \theta_{(m)} \sigma_{(m)} \beta_{(m)}^T \\
&\approx \theta \, \sigma \, \beta^T && \text{(Rank 1 approx.)}
\end{aligned}
$$

σ says *how much relative emphasizing* is happening in this dimension

Infer dimensional structure

Intuition: λ has a orthogonal decomposition

$$\begin{aligned} \lambda &= \Theta \Sigma B^T && \text{(SVD)} \\ &= \sum_m^M \theta_{(m)} \sigma_{(m)} \beta_{(m)}^T \\ &\approx \theta \, \sigma \, \beta^T && \text{(Rank 1 approx.)} \end{aligned}$$

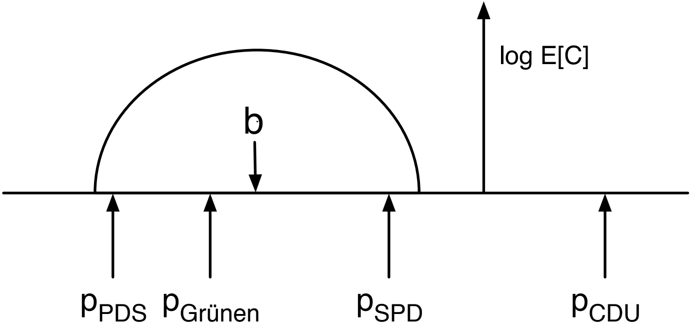
σ says *how much relative emphasizing* is happening in this dimension

Right now, there's only one dimension so it's not so interesting…

Intuition

What are we doing when we fit such a model?

A generative model of positioning text



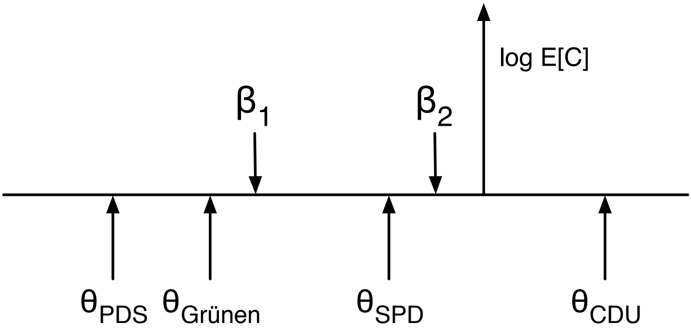
$$\log \mu_{ij} = r_i + c_j \ + \ \frac{(p_i - b_j)^2}{v}$$

This is just our model with a false moustache and hat

Quadratic unfolding (Elff 2013, Heiser 1986) has the model as a reduced form

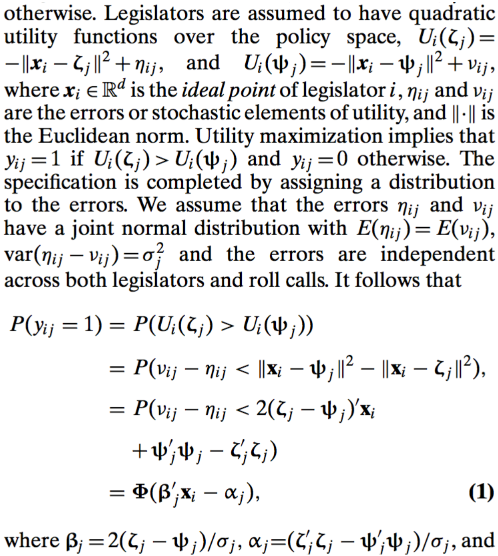
$$
\begin{aligned}
\log \mu_{ij} &= r_i + c_j \; + \; \frac{(p_i - b_j)^2}{v} \\
&= r_i + c_j + (p_i^2 - 2p_i b_j + b_j^2)/v \\
&= [r_i + p_i^2/v] + [c_j + b_j^2/v] + [p_i]\,[1/v][-2b_j] \\
&= \quad a_i \qquad + \qquad \psi_j \quad + \; \theta_i \quad \sigma \quad\; \beta_j
\end{aligned}
$$

Just like spatial voting



Two words/topics, e.g. 'benefits' and 'assets', with scores β_1 and β_2 in a document of length N_i

Just like spatial voting



otherwise. Legislators are assumed to have quadratic utility functions over the policy space, $U_i(\zeta_j) = -\|\boldsymbol{x}_i - \zeta_j\|^2 + \eta_{ij}$, and $U_i(\psi_j) = -\|\boldsymbol{x}_i - \psi_j\|^2 + \nu_{ij}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the *ideal point* of legislator i, η_{ij} and ν_{ij} are the errors or stochastic elements of utility, and $\|\cdot\|$ is the Euclidean norm. Utility maximization implies that $y_{ij} = 1$ if $U_i(\zeta_j) > U_i(\psi_j)$ and $y_{ij} = 0$ otherwise. The specification is completed by assigning a distribution to the errors. We assume that the errors η_{ij} and ν_{ij} have a joint normal distribution with $E(\eta_{ij}) = E(\nu_{ij})$, $\text{var}(\eta_{ij} - \nu_{ij}) = \sigma_j^2$ and the errors are independent across both legislators and roll calls. It follows that

$$\begin{aligned} P(y_{ij} = 1) &= P(U_i(\zeta_j) > U_i(\psi_j)) \\ &= P(\nu_{ij} - \eta_{ij} < \|\mathbf{x}_i - \psi_j\|^2 - \|\mathbf{x}_i - \zeta_j\|^2), \\ &= P(\nu_{ij} - \eta_{ij} < 2(\zeta_j - \psi_j)'\mathbf{x}_i \\ &\quad + \psi_j'\psi_j - \zeta_j'\zeta_j) \\ &= \Phi(\beta_j'\mathbf{x}_i - \alpha_j), \qquad \textbf{(1)} \end{aligned}$$

where $\beta_j = 2(\zeta_j - \psi_j)/\sigma_j$, $\alpha_j = (\zeta_j'\zeta_j - \psi_j'\psi_j)/\sigma_j$, and

From Clinton et al. (2004)

Just like spatial voting

$$
\begin{aligned}
[C_{i1}, C_{i2}] &\sim \text{Binomial}([\pi_{i1}, \pi_{i2}], N_i) \\
\pi_{i1} &= \mu_{i1}/(\mu_{i1} + \mu_{i2}) \\
\log\left(\frac{\pi_{i1}}{\pi_{i2}}\right) &= \log \pi_{i1} - \log \pi_{i2} \\
&= (a_i - a_i) + (\psi_1 - \psi_2) + \theta_i\,(\beta_1 - \beta_2) \\
&= \qquad\qquad\qquad\quad \psi_{1/2} \quad + \theta_i \quad \beta_{1/2}
\end{aligned}
$$

Just like spatial voting

$$
\begin{aligned}
[C_{i1}, C_{i2}] &\sim \text{Binomial}([\pi_{i1}, \pi_{i2}], N_i) \\
\pi_{i1} &= \mu_{i1} / (\mu_{i1} + \mu_{i2}) \\
\log\left(\frac{\pi_{i1}}{\pi_{i2}}\right) &= \log \pi_{i1} - \log \pi_{i2} \\
&= (\alpha_i - \alpha_i) + (\psi_1 - \psi_2) + \theta_i \, (\beta_1 - \beta_2) \\
&= \qquad\qquad\qquad\quad \psi_{1/2} \quad + \theta_i \quad \beta_{1/2}
\end{aligned}
$$

Look Ma, a logit!

Special case: logit scores

Identify left L and right R topic and compute (Lowe et al. 2011)

$$\hat{\theta}_i = \log \frac{\sum_{j \in R} C_{ij}}{\sum_{k \in L} C_{ik}}$$

Special case: logit scores

Identify left L and right R topic and compute (Lowe et al. 2011)

$$\hat{\theta}_i = \log \frac{\sum_{j \in R} C_{ij}}{\sum_{k \in L} C_{ik}}$$

Position is relative *proportional* emphasis, with a psychophysical motivation

Revisiting human judgement

(Budge et al. 1983, Baumgartner and Jones)

Validating what comes out of the smoky room

The CMP project have performed a huge manual content analysis and *chosen* some right and left topics for us.

> This kind of thing is a popular exercise (unless you're the coder)

We're supposed to add both sides up and subtract to get a position measure for documents

Validating what comes out of the smoky room

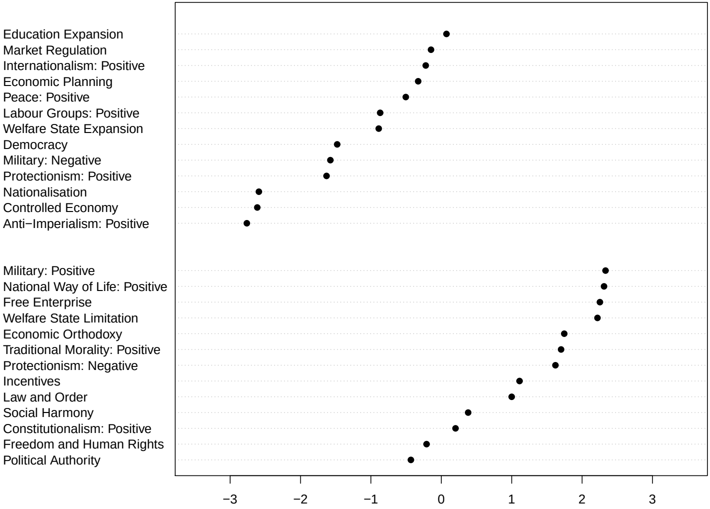
The CMP project have performed a huge manual content analysis and *chosen* some right and left topics for us.

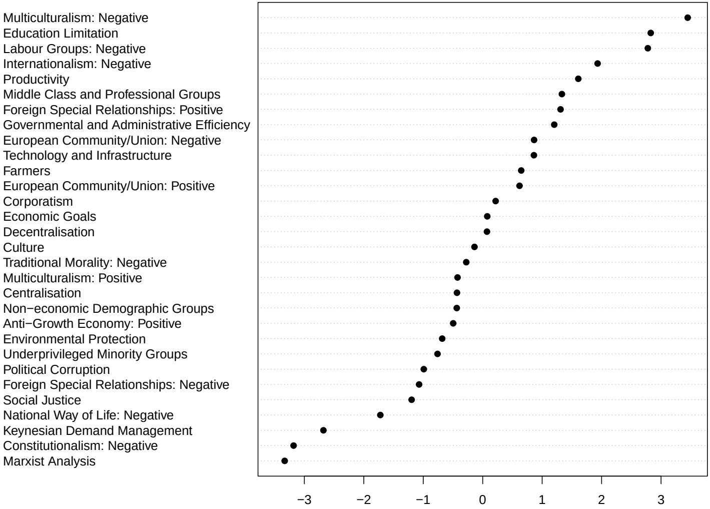
- This kind of thing is a popular exercise (unless you're the coder)

We're supposed to add both sides up and subtract to get a position measure for documents

Are these topics really used by parties on right and left?

- Let's run our model on the topic output and check
- Turns out our $\hat{\theta}$s correlate 0.94 with their scale
- We're more interested in βs


Education Expansion
Market Regulation
Internationalism: Positive
Economic Planning
Peace: Positive
Labour Groups: Positive
Welfare State Expansion
Democracy
Military: Negative
Protectionism: Positive
Nationalisation
Controlled Economy
Anti−Imperialism: Positive
Military: Positive
National Way of Life: Positive
Free Enterprise
Welfare State Limitation
Economic Orthodoxy
Traditional Morality: Positive
Protectionism: Negative
Incentives
Law and Order
Social Harmony
Constitutionalism: Positive
Freedom and Human Rights
Political Authority
−3
−2
−1
0
1
2
3


Multiculturalism: Negative
Education Limitation
Labour Groups: Negative
Internationalism: Negative
Productivity
Middle Class and Professional Groups
Foreign Special Relationships: Positive
Governmental and Administrative Efficiency
European Community/Union: Negative
Technology and Infrastructure
Farmers
European Community/Union: Positive
Corporatism
Economic Goals
Decentralisation
Culture
Traditional Morality: Negative
Multiculturalism: Positive
Centralisation
Non−economic Demographic Groups
Anti−Growth Economy: Positive
Environmental Protection
Underprivileged Minority Groups
Political Corruption
Foreign Special Relationships: Negative
Social Justice
National Way of Life: Negative
Keynesian Demand Management
Constitutionalism: Negative
Marxist Analysis
−3
−2
−1
0
1
2
3

Dimension issues

What the heck is θ?

How can we be sure that there is only one of them?

What is θ?

Whatever maximizes the Likelihood…

What is θ?

Whatever maximizes the Likelihood…

Like all scaling techniques (e.g. NOMINATE), this model is *exploratory* – *you* have to figure out what the dimension really is.

One dimensional world

How do we know that positions on only one dimension are being expressed?

Relatedly: how do we get positions on a specific policy issue?

One dimensional world

How do we know that positions on only one dimension are being expressed?

Relatedly: how do we get positions on a specific policy issue?

Three possibilities

- Use only those texts (or sections thereof) that are guaranteed to be on the same topic and *scale them separately* (Slapin and Proksch, 2008)
- Learn items from just a subset of relevant documents (Laver et al. 2003)
- Work with *topic* counts rather than word counts (Baerg and Lowe, MS)

Heroic assumptions are (closer to being) true

Multidimensional world

Allow for more dimensions! $\theta_i^1, \theta_i^2, \ldots$

We need to move to a computationally cheaper model:

Correspondence analysis (Greenacre 2007)

For identification, a K-dimensional model has K sets of θ and K sets of β

and they'll be orthogonal…

Multidimensional world

Allow for more dimensions! $\theta_i^1, \theta_i^2, \ldots$

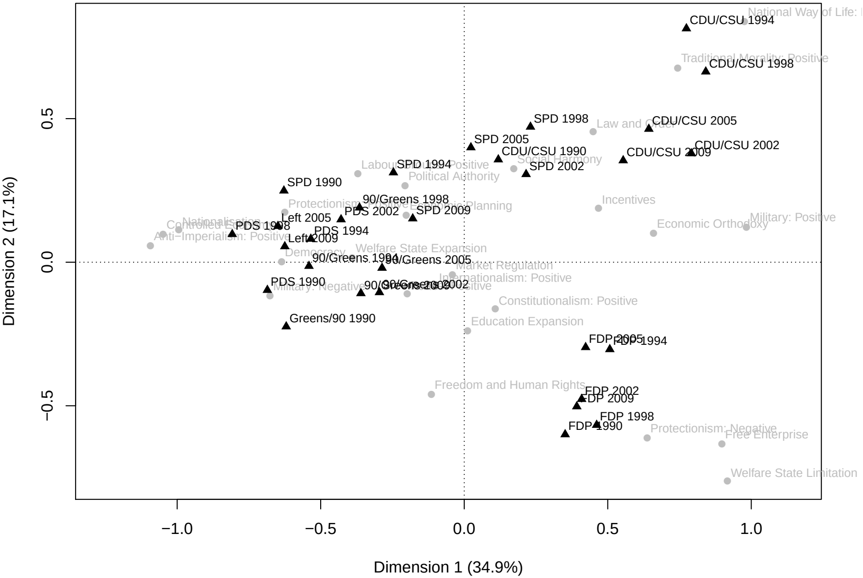
We need to move to a computationally cheaper model:

Correspondence analysis (Greenacre 2007)

For identification, a K-dimensional model has K sets of θ and K sets of β

and they'll be orthogonal…

Fit this model to the German topic counts…


National Way of Life:
CDU/CSU 1994
Traditional Morality: Positive
CDU/CSU 1998
SPD 1998
Law and Order
CDU/CSU 2005
SPD 2005
CDU/CSU 1990
CDU/CSU 2002
CDU/CSU 2009
Social Harmony
SPD 2002
Labour Groups: Positive
SPD 1994
Political Authority
SPD 1990
Incentives
90/Greens 1998
Protectionism: Positive
Economic Planning
PDS 2002
SPD 2009
Military: Positive
Economic Orthodoxy
Left 2005
Nationalisation
Controlled Economy
PDS 1998
PDS 1994
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
Positive
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
Dimension 2 (17.1%)
−1.0
−0.5
0.0
0.5
1.0
Dimension 1 (34.9%)

(Graduate student) life skills

How to read a biplot:

- Documents points are closer when using words/topics *similarly*
- Words points are closer with *similar* document profiles
- 0,0: a document or word/topic used *exactly as often as we would expect by chance*
- Document vector: arrow from 0,0 to a document point
- Word/topic vector: arrow from 0,0 to a word/topic point
- Vectors are *longer* the more their usage diverges from chance
- *Angle* between a word vector and document vector: how much a document preferentially uses the word

(Graduate student) life skills

There is nothing special to text about a biplot

This interpretation works for *all kinds* of cross-tables.

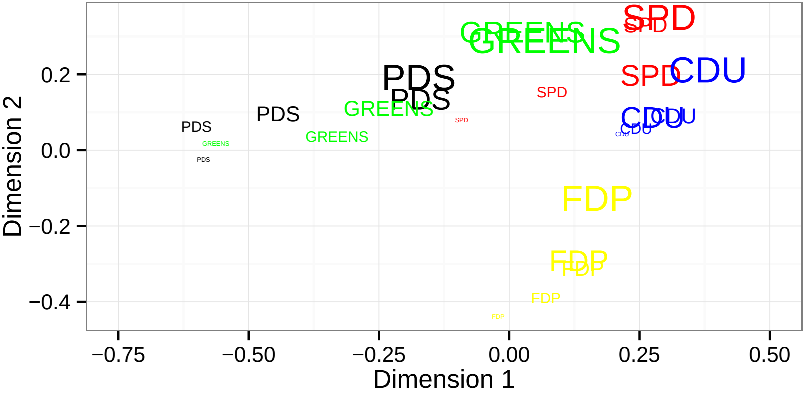
Use it for good!

Dimensions and topic change

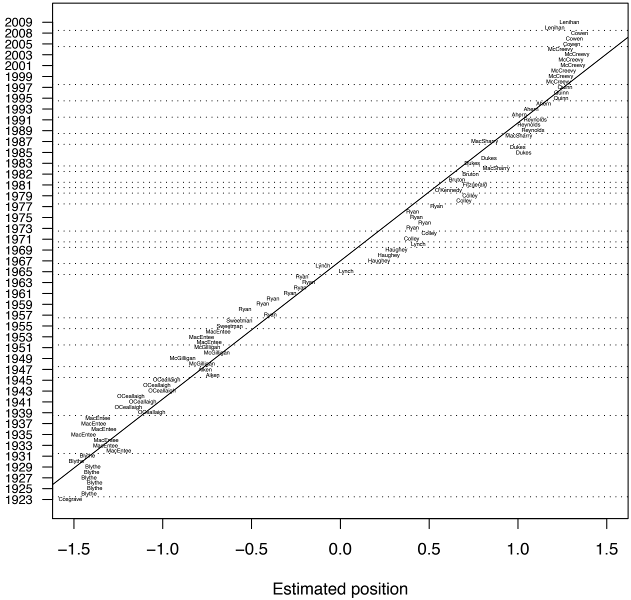
What if the political lexicon changes over time? (it does)

New issues appear, old issues disappear

Then scaling algorithms pick up shifts in the policy agenda rather than shifts in party positions.


SPD
SPD
GREENS
GREENS
CDU
PDS
SPD
PDS
SPD
GREENS
PDS
SPD
CDU
CDU
CDU
CDU
PDS
GREENS
GREENS
PDS
FDP
FDP
FDP
FDP
FDP
Dimension 2
0.2
0.0
−0.2
−0.4
−0.75
−0.50
−0.25
0.00
0.25
0.50
Dimension 1

Worst Case Scenario



Lab Time