(Un)Supervised scaling of texts

Mate Akos

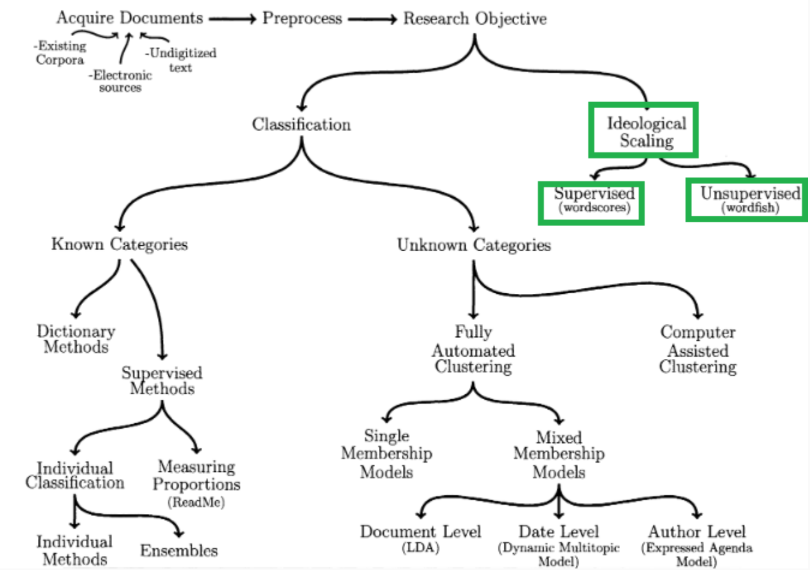
2019 November

TK PTI

For today

1. Where we are
2. Scaling texts
3. Supervised: Wordscores
4. Unsupervised: Wordfish

Where we are


Acquire Documents
Preprocess
Research Objective
-Existing Corpora
-Electronic sources
-Undigitized text
Classification
Ideological Scaling
Supervised (wordscores)
Unsupervised (wordfish)
Known Categories
Unknown Categories
Dictionary Methods
Supervised Methods
Fully Automated Clustering
Computer Assisted Clustering
Individual Classification
Measuring Proportions (ReadMe)
Single Membership Models
Mixed Membership Models
Individual Methods
Ensembles
Document Level (LDA)
Date Level (Dynamic Multitopic Model)
Author Level (Expressed Agenda Model)

Scaling

Scaling texts:

- Key interest of social science. We are interested in the latent dimension(s) of the texts
    1. Political scales (left-right)
    2. Policy positions
    3. Any choosen continuous scale
- Large human coded projects:

Scaling

Scaling texts:

- Key interest of social science. We are interested in the latent dimension(s) of the texts
    1. Political scales (left-right)
    2. Policy positions
    3. Any choosen continuous scale
- Large human coded projects:
    1. Manifesto Project (political manifesto coding)
    2. Chapel Hill expert survey

Scaling

Problems with human coding

- Expensive, resource intensive

Scaling

Problems with human coding

- Expensive, resource intensive
- Each text needs to be coded by 2 coder

Scaling

Problems with human coding

- Expensive, resource intensive
- Each text needs to be coded by 2 coder
- inter-coder reliability tends to be low

Scaling

Problems with human coding

- Expensive, resource intensive
- Each text needs to be coded by 2 coder
- inter-coder reliability tends to be low

Solution:

Scaling

Problems with human coding

- Expensive, resource intensive
- Each text needs to be coded by 2 coder
- inter-coder reliability tends to be low

Solution:

- supervised: Wordscores
- unsupervised: Wordfish

Wordscores

From Laver, Benoit and Garry (2003): Extracting Policy Positions from Political Texts Using Words as Data

- **Reference texts**: documents scored already on a scale
- **Virgin texts**: documents without any information
- Similar concept to the training and test sets introduced in the classification session
- Core idea:
    1. Get word scores from reference texts
    2. Apply those word scores on the virgin texts

Wordscores

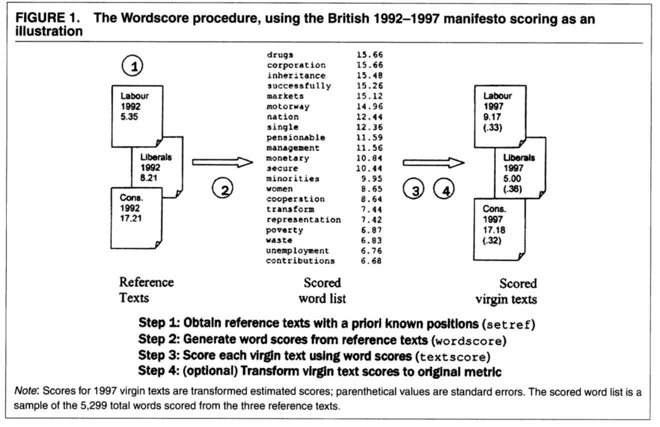


FIGURE 1. The Wordscore procedure, using the British 1992–1997 manifesto scoring as an illustration

drugs	15.66
corporation	15.66
inheritance	15.48
successfully	15.26
markets	15.12
motorway	14.96
nation	12.44
single	12.36
pensionable	11.59
management	11.56
monetary	10.84
secure	10.44
minorities	9.95
women	8.65
cooperation	8.64
transform	7.44
representation	7.42
poverty	6.87
waste	6.83
unemployment	6.76
contributions	6.68

Step 1: Obtain reference texts with a priori known positions (`setref`)
Step 2: Generate word scores from reference texts (`wordscore`)
Step 3: Score each virgin text using word scores (`textscore`)
Step 4: (optional) Transform virgin text scores to original metric

Note: Scores for 1997 virgin texts are transformed estimated scores; parenthetical values are standard errors. The scored word list is a sample of the 5,299 total words scored from the three reference texts.

Wordscores - Reference texts

Used as benchmarks

- We need information on their positions on the scale
- The more reference texts we have the better

Get the word scores

(1) Probability of having world w in our reference text r (F_{wr} is frequency of word w in reference text r):

$$P_{wr} = \frac{F_{wr}}{\sum_r F_{wr}}$$

(2) Assigning word scores for each word w for each dimension d (A_{rd} is the known position of reference text r on dimension d)

$$S_{wd} = \sum_r P_{wr} A_{rd}$$

Get the word scores

(3) We score each virgin text *v* on dimension *d* using the wordscore S_{wd} (F_{wv}) is the frequency of words *w* in virgin text *v*

$$S_{vd} = \sum_{w} F_{wv} S_{w} d$$

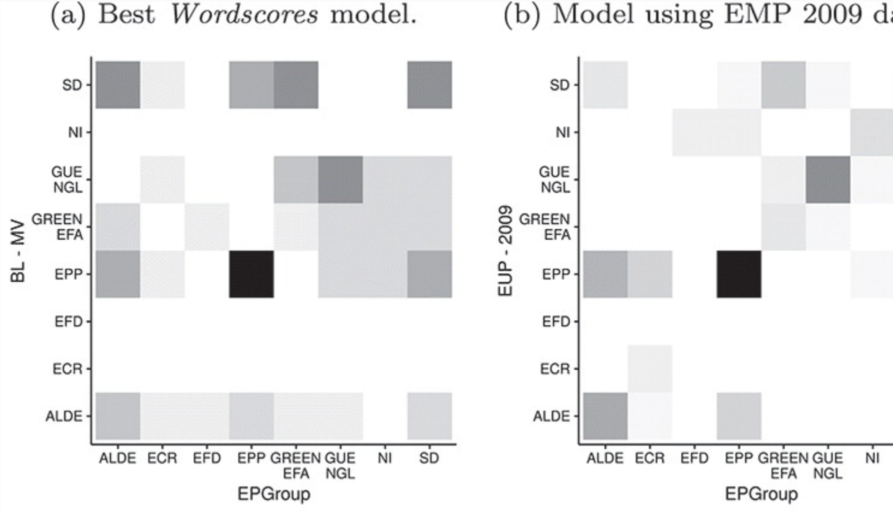
(4) Rescaling raw text scores

$$S^{*}_{vd} = (S_{vd} - S_{\bar{vd}})(\frac{SD_{rd}}{SD_{vd}}) + S_{\bar{vd}}$$

How to select reference texts

1. Results depends on reference text selection
2. Reference text scores need to be validated
3. Stemming decreases effectiveness (Ruedin 2013)
4. Dimensions must be known of the reference texts a priori
5. reference texts should come from the extremes to provide anchors
6. The reference and virgin texts should have the same frame of reference (cannot score movie reviews based on political manifestos)
7. High word count!

Wordscores



(a) Best *Wordscores* model.

(b) Model using EMP 2009 da

Wordfish

Slapin & Proksch, 2008: A Scaling Model for Estimating Time-Series Party Positions from Texts

- unsupervised scaling of **political** texts
- does not need reference documents
- Has only one parameter: λ (intensity parameter)
- Word frequencies are generated by a Poission process
  - Poisson is a discrete probability distribution which models the lenght of time waiting for some event if the probability of the event is proportional to the length of the wait.

Functional form

$$y_{ijt} \sim \text{Poisson}(\lambda_{ijt})$$

$$\lambda_{ijt} = exp(\alpha_{it} + \psi_j + \beta_j * \theta_{it})$$

Where

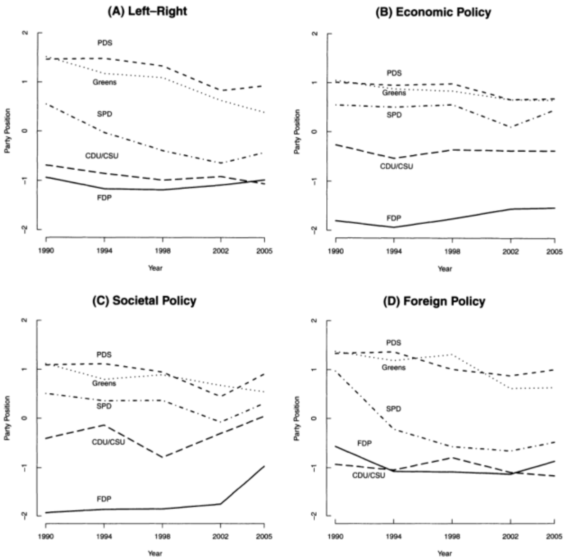
- y_{ijt} count of word *j* in party manifesto *i* in year *t*.
- α_{it} is the party-year fixed effects (accounting for the differences in manifesto lengths)
- ψ_j word fixed effects (allowing for differing word frequencies)
- β_j word weight, capturing the importance of word *j* in discriminating party positions
- θ_{it} estimate of party *i*'s position in year *t*

Estimating the Wordfish model

1. Calculate starting values
2. Estimate party parameters
3. Estimate word parameters
4. Calculate log-likelihood
5. Repeat steps 2-4 until convergence (change in values is below a certain threshold)

Wordscores

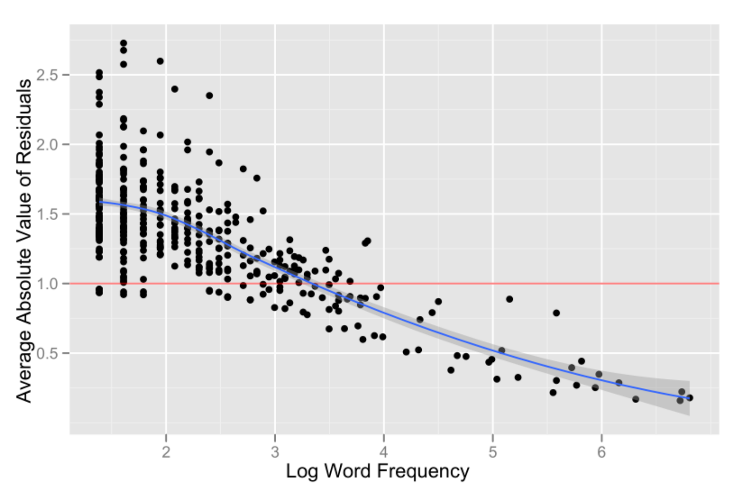
FIGURE 1 Estimated Party Positions in Germany, 1990–2005



Problems

- Conditional independence does not hold
  - words are in sequence (serial correlation)
  - certain words occur in combination (e.g.: ”vice president”)
- Heteroskedasticity of the errors
  - when informative words cluster -> overdispersion
  - when high frequency words are uninformative -> underdispersion

Overdispersion



source: Pablo Barberá, LSE `https://github.com/lse-my459/lectures/tree/master/week07`

Validating Wordscores and Wordfish

next week.