12: Tekst som data

Videregående kvantitative metoder i studiet af politisk adfærd

Frederik Hjorth
fh@ifs.ku.dk
fghjorth.github.io
@fghjorth

Institut for Statskundskab
Københavns Universitet

29. november 2017

1. Formalia
2. Opsamling fra sidst
3. Intro til text as data
4. Klassifikation
5. Skalering
6. Case: Baturo & Mikhaylov
7. Kig fremad

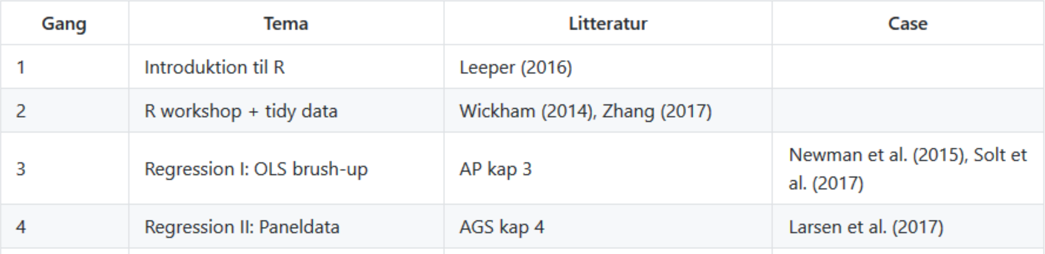
Eksamen

- frist for seminaropgave: 22. december 23.59
- 1. genindlevering: 2. januar 23.59
- 2. genindlevering: 5. januar 23.59

→ alt afleveres på Absalon

Fagets opbygning

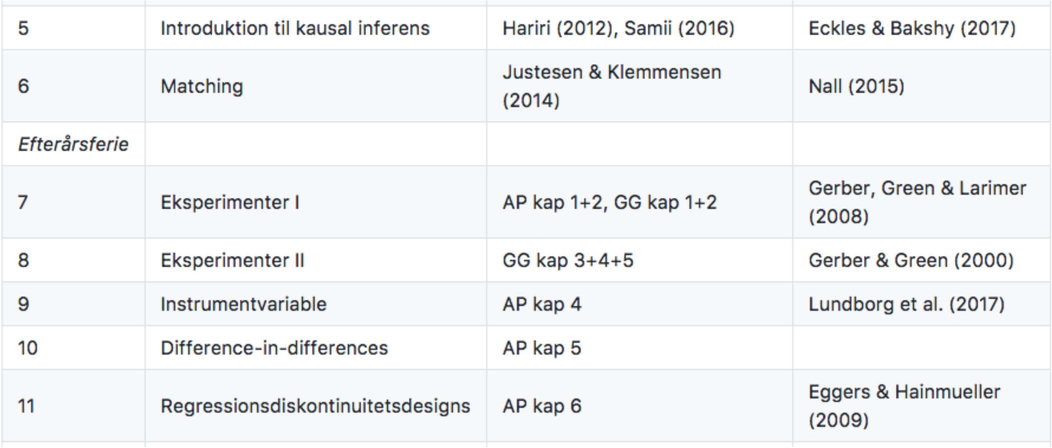
Blok 1



Gang	Tema	Litteratur	Case
1	Introduktion til R	Leeper (2016)	
2	R workshop + tidy data	Wickham (2014), Zhang (2017)	
3	Regression I: OLS brush-up	AP kap 3	Newman et al. (2015), Solt et al. (2017)
4	Regression II: Paneldata	AGS kap 4	Larsen et al. (2017)

Fagets opbygning

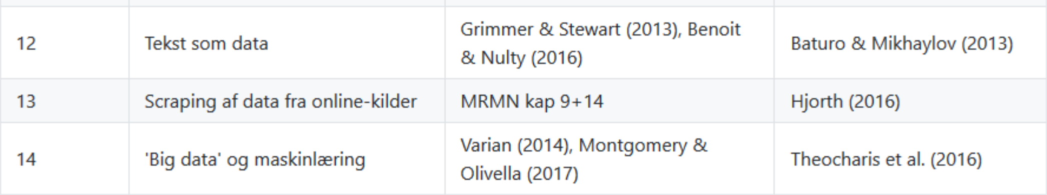
Blok 2



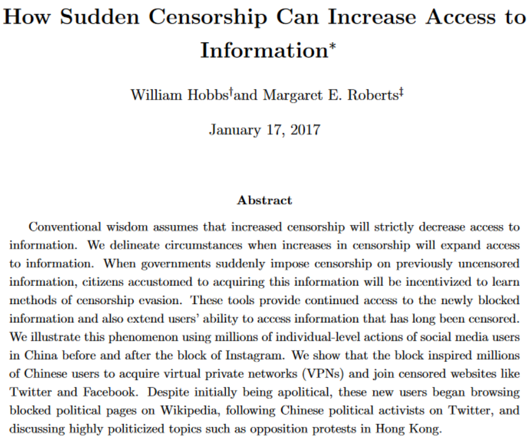
5	Introduktion til kausal inferens	Hariri (2012), Samii (2016)	Eckles & Bakshy (2017)
6	Matching	Justesen & Klemmensen (2014)	Nall (2015)
Efterårsferie			
7	Eksperimenter I	AP kap 1+2, GG kap 1+2	Gerber, Green & Larimer (2008)
8	Eksperimenter II	GG kap 3+4+5	Gerber & Green (2000)
9	Instrumentvariable	AP kap 4	Lundborg et al. (2017)
10	Difference-in-differences	AP kap 5	
11	Regressionsdiskontinuitetsdesigns	AP kap 6	Eggers & Hainmueller (2009)

Fagets opbygning

Blok 3



12	Tekst som data	Grimmer & Stewart (2013), Benoit & Nulty (2016)	Baturo & Mikhaylov (2013)
13	Scraping af data fra online-kilder	MRMN kap 9+14	Hjorth (2016)
14	'Big data' og maskinlæring	Varian (2014), Montgomery & Olivella (2017)	Theocharis et al. (2016)



How Sudden Censorship Can Increase Access to Information*

William Hobbs[†] and Margaret E. Roberts[‡]

January 17, 2017

Abstract

Conventional wisdom assumes that increased censorship will strictly decrease access to information. We delineate circumstances when increases in censorship will expand access to information. When governments suddenly impose censorship on previously uncensored information, citizens accustomed to acquiring this information will be incentivized to learn methods of censorship evasion. These tools provide continued access to the newly blocked information and also extend users' ability to access information that has long been censored. We illustrate this phenomenon using millions of individual-level actions of social media users in China before and after the block of Instagram. We show that the block inspired millions of Chinese users to acquire virtual private networks (VPNs) and join censored websites like Twitter and Facebook. Despite initially being apolitical, these new users began browsing blocked political pages on Wikipedia, following Chinese political activists on Twitter, and discussing highly politicized topics such as opposition protests in Hong Kong.

Talk m. Molly Roberts **onsdag d. 6. december 10-11 i frokoststuen** `http://bit.ly/vkme17evaluering`

Opsamling fra sidst

- eksempel: National Merit Award Scholarships
- logikken i RD
- formel definition
- RD i regressionsform
- udfordringer: funktionel form, båndbredde, sorting
- RD vs. diff-in-diff
- case: Eggers & Hainmueller

Udgangspunkt: mange politisk relevante fænomener er tekstlige + stor del af 'data-revolutionen' udgøres af tekstdata

- folketingsdebatter
- nytårstaler
- partiprogrammer
- regeringsprogrammer
- udvalgsspørgsmål
- fritekstsvar i kandidattests
- politikeres emails
- — "— facebook-opdateringer
- — "— tweets
- etc. etc.

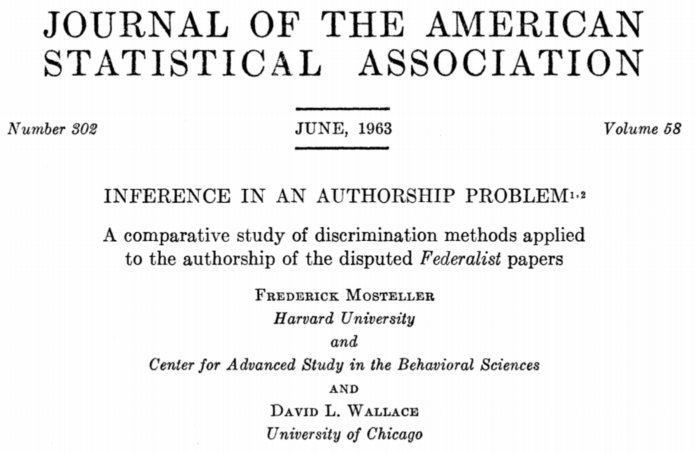
→ behov for metoder til at overskue/analysere data

Ex.:

> The accumulation of all powers, legislative, executive, and judiciary, in the same hands, whether of one, a few, or many, and whether hereditary, self-appointed, or elective, may justly be pronounced the very definition of tyranny.

> Udgangspunktet for regeringen er VK-regeringens økonomiske politik i bredeste forstand, herunder genopretningsaftalen og forårets aftaler herunder tilbagetrækningsreformen. Regeringen vil gennemføre reformer, der øger arbejdsudbuddet, så vi kan øge væksten i dansk økonomi, sikre holdbare offentlige finanser, og en beskeden og målrettet udbygning af den offentlige service.

Pioner-studie: Mosteller & Wallace om *Federalist Papers*



JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

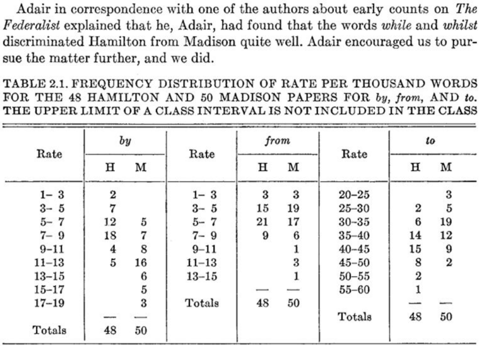
Number 302 JUNE, 1963 *Volume 58*

INFERENCE IN AN AUTHORSHIP PROBLEM[1,2]

A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
DAVID L. WALLACE
University of Chicago

Pioner-studie: Mosteller & Wallace om *Federalist Papers*



Adair in correspondence with one of the authors about early counts on *The Federalist* explained that he, Adair, had found that the words *while* and *whilst* discriminated Hamilton from Madison quite well. Adair encouraged us to pursue the matter further, and we did.

TABLE 2.1. FREQUENCY DISTRIBUTION OF RATE PER THOUSAND WORDS FOR THE 48 HAMILTON AND 50 MADISON PAPERS FOR *by*, *from*, AND *to*. THE UPPER LIMIT OF A CLASS INTERVAL IS NOT INCLUDED IN THE CLASS

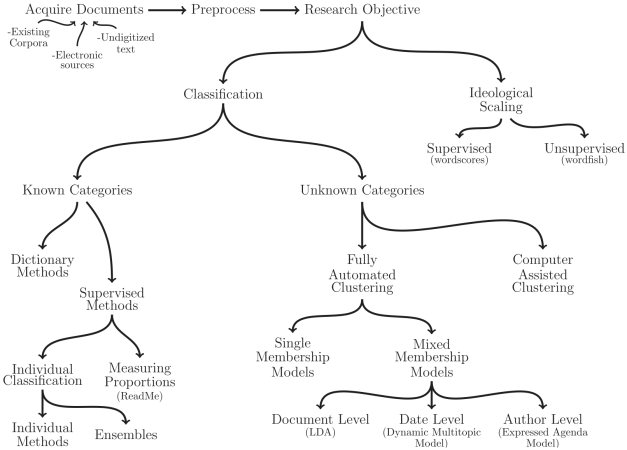
Rate	*by*		Rate	*from*		Rate	*to*	
	H	M		H	M		H	M
1– 3	2		1– 3	3	3	20–25		3
3– 5	7		3– 5	15	19	25–30	2	5
5– 7	12	5	5– 7	21	17	30–35	6	19
7– 9	18	7	7– 9	9	6	35–40	14	12
9–11	4	8	9–11		1	40–45	15	9
11–13	5	16	11–13		3	45–50	8	2
13–15		6	13–15		1	50–55	2	
15–17		5		—	—	55–60	1	
17–19		3	Totals	48	50		—	—
	—	—				Totals	48	50
Totals	48	50						

Frederick Mosteller, Harvard University

Source: Mosteller, Wallace, *Inference in an authorship problem: A comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers*, Journal of the American Statistical Association, Volume 58, issue 302, 1963.

Overordnet sondring:

- klassifikation → hvad handler teksterne om? (kategorisk outcome)
- skalering → hvordan er teksterne fordelt på en skala? (kontinuert outcome)



Sondring inden for både klassifikation og skalering:

- superviseret: tekster klassificeres/skaleres pba. udvalgte tekster med ‘kendte’ værdier
- usuperviseret: tekster klassificeres alene pba. data i teksterne

central forskel: menneskelig fortolkning før estimation (superviseret) eller efter (usuperviseret)
→ denne sondring vender tilbage om 14 dage!

- udgangspunkt for næsten al text as data: *bag-of-words assumption*
- m.a.o.: teksters betydning afspejles i ordfrekvenser
- men antager også at ordrækkefølge er irrelevant
- oplagte modeks., fx. `mindre stat, mere privat` ctr. `mere stat, mindre privat`
- rækkefølge kan principielt håndteres m. bigrams, trigrams, ... n-grams
- men: n-grams computationelt bekosteligt, generelt beskeden analytisk gevinst

Grimmer & Stewart: fire principper for tekstanalyse

1. alle modeller er forkerte, men nogle er brugbare
2. kvantitative tekstanalysemetoder understøtter menneskelig læsning
3. der findes ikke én globalt optimal metode
4. validér, validér, validér

Typisk proces for tekstanalyse i dag:

1. import af tekster som et *korpus*
2. pre-processering:
   - fjern tal, specialtegn
   - fjern 'stopwords'
   - stemming
   - fjern meget sjældne el. hyppige ord
3. konvertering til *document-term*/*document-feature* matrice
4. analyse

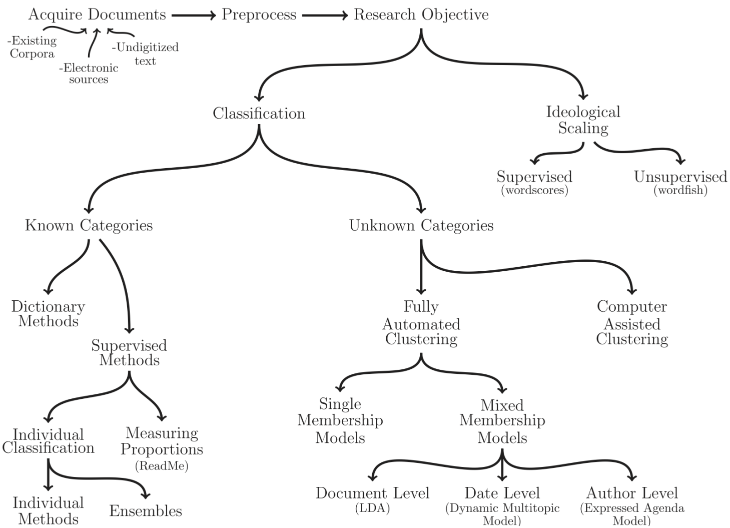
Eks. på document-term-matrice:

```
## inspect first 5 rows and first 8 columns
inspect(dtm[1:5, 1:8])

## <<DocumentTermMatrix (documents: 5, terms: 8)>>
## Non-/sparse entries: 4/36
## Sparsity           : 90%
## Maximal term length: 7
## Weighting          : term frequency (tf)
##
##           Terms
## Docs        abandon abat abb abet abhorr abil abject abl
##   fp01.txt        0    0   0    0      0    0      0   1
##   fp02.txt        0    0   0    0      0    1      0   0
##   fp03.txt        0    0   0    0      0    0      0   2
##   fp04.txt        0    0   0    0      0    0      0   1
##   fp05.txt        0    0   0    0      0    0      0   0
```

- klassisk pakke til text as data: `tm`
- nyere, enklere alternativ: `quanteda` af Ken Benoit et al.
- fremgangsmåde m. `quanteda`:
    1. import m. `readtext()` i standalone-pakken `readtext`
    2. definition som korpus m. `corpus()`
    3. preprocessering+konvertering m. `dfm()`
    4. analyse, fx. m. `textmodel_*()`

→ vi gennemgår dette i casen!



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

Individual Classification

Measuring Proportions (ReadMe)

Individual Methods

Ensembles

Fully Automated Clustering

Computer Assisted Clustering

Single Membership Models

Mixed Membership Models

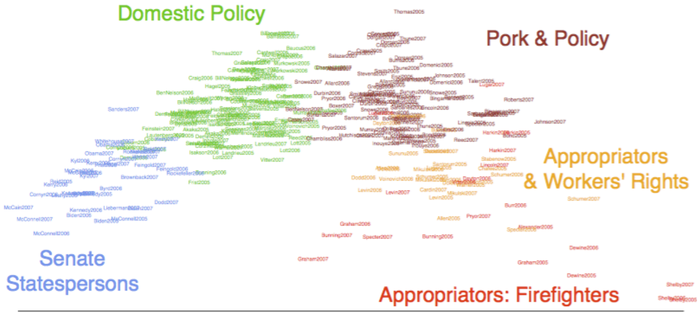
Document Level (LDA)

Date Level (Dynamic Multitopic Model)

Author Level (Expressed Agenda Model)

- hvad handler teksterne om?
- ⇝ hvilke latente kategorier (emner) udspringer teksterne af?
- typisk anvendt approach: emnemodeller (topic models)
- her: *tf-idf* → ret primitiv, men letforståelig

Grimmer (2013): Analyse af 64k pressemeddelelser



Senate Position Taker

- Iraq War
- Intelligence
- Intl. Relations
- Budget

Domestic Policy

- Environment
- Gas prices
- DHS
- Consumer Safety

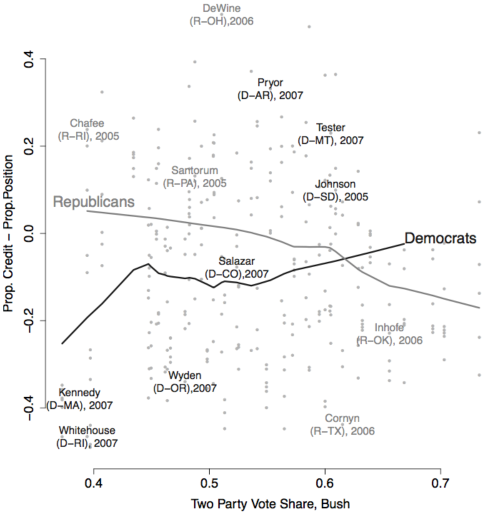
Pork & Policy

- WRDA grants
- Farming
- Health Care
- Education Policy

Appropriators

- Fire Grants
- Airport Grants
- University Money
- Police Grants

Grimmer (2013): Analyse af 64k pressemeddelelser



term frequency for term t i dokument d:

$$tf = f_{td}$$

inverse document frequency:

$$idf = log\left(\frac{N}{n_t}\right)$$

term frequency-inverse document frequency (tf-idf):

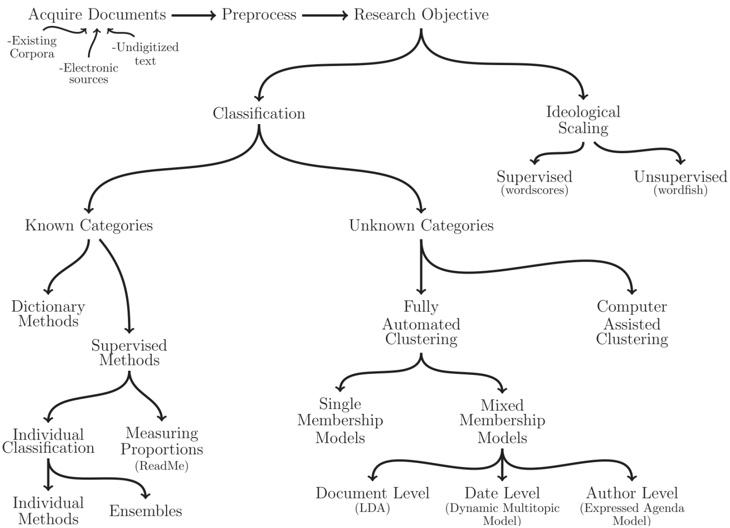
$$tf \times idf = f_{td} \times log\left(\frac{N}{n_t}\right)$$

Fire stiliserede partiprogrammer:

parti	partiprogram
Enh.	velfærd velfærd velfærd
S	velfærd velfærd vækst
V	velfærd vækst vækst
LA	vækst vækst vækst

→ hvad er tf-idf for 'velfærd' hos Enhedslisten?

$$tf \times idf = f_{td} \times log\left(\frac{N}{n_t}\right)$$


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)

For dokumentet d med W ordtyper ('tokens') estimerer vi positionen θ_d:

$$\hat{\theta}_d = \frac{1}{W} \sum_{w=1}^{W} \hat{\pi}_w \tag{1}$$

for R referencetekster estimeres $\hat{\pi}_w$:

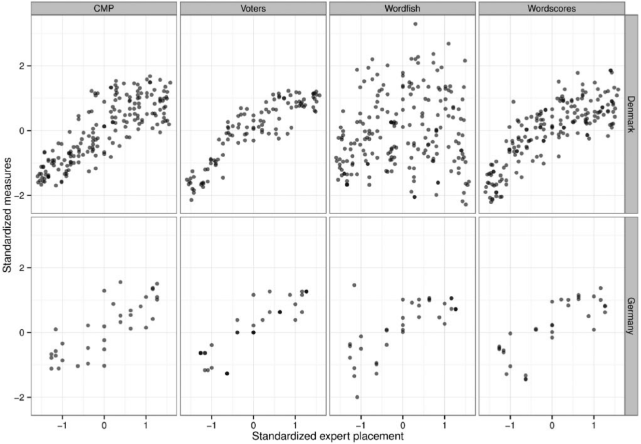
$$\hat{\pi}_w = \sum_{r=1}^{R} \theta_r \hat{P}(d_r | w) \tag{2}$$

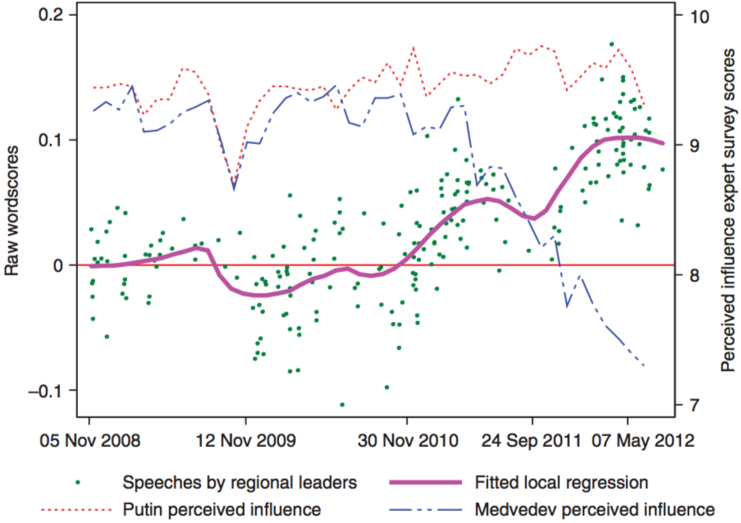
hvor pr. Bayes' teorem:

$$\hat{P}(d_r|w) = \frac{\hat{P}(w|d_i)}{\sum_{r=1}^{R} \hat{P}(w|d_r)} \tag{3}$$

$\rightarrow$ wordscoren $\hat{\pi}_w$ sammenvejer hvert ref-tekst r's position med hvor stærkt d prædikerer r

Hjorth et al. (2015): Wordscores reproducerer ekspertestimater af partiprogrammer (men alternativet Wordfish gør ikke)




0.2
0.1
0
−0.1
Raw wordscores
10
9
8
7
Perceived influence expert survey scores
05 Nov 2008
12 Nov 2009
30 Nov 2010
24 Sep 2011
07 May 2012
Speeches by regional leaders
Fitted local regression
Putin perceived influence
Medvedev perceived influence

Næste gang: data fra nettet

- screen scraping
- API'er
- pensum: MRMN kap. 9+14
- vigtigt i kap. 9: 9.1.10+
- case: Hjorth (ananas i egen juice ⇝ eksempel på data fra online-kilder, læs kursorisk)
- ekstra hjemmearbejde: `lav en twitter API key`

Tak for i dag!