4: 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

27. september 2018

Opsamling fra sidst

- online data
- web scraping
- etik i web scraping
- API'er
- case I: Hjorth (2016)
- case II: skalering af danske Twitter-brugere

- OBS: ingen undervisning torsdag d. 4. oktober
- erstatning: onsdag d. 10. oktober 15-17 i lokale 7.0.18

Fagets opbygning

Gang	Tema	Litteratur	Case
1	Introduktion til R	Leeper (2016)	
2	R workshop I + tidy data	Wickham (2014), Zhang (2017)	
3	Data fra online-kilder	MRMN kap 9+14	Hjorth (2016)
4	Tekst som data	Grimmer & Stewart (2013), Benoit & Nulty (2016)	Baturo & Mikhaylov (2013)
5	Regression I: OLS brush-up	AP kap 3	Mutz (2018)
6	Regression II: Paneldata	AGS kap 4	Mutz (2018)
Efterårsferie			

Fagets opbygning

7	R workshop II	tba	
8	Introduktion til kausal inferens	Angrist & Pischke (2010), Samii (2016)	Carroll (2018)
9	Eksperimenter I	AP kap 1+2, GG kap 1+2	Gerber, Green & Larimer (2008)
10	Eksperimenter II	GG kap 3+4+5	Gerber & Green (2000)
11	Instrumentvariable	AP kap 4	Colantone & Stanig (2018)
12	Difference-in-differences	AP kap 5	
13	Regressionsdiskontinuitetsdesigns	AP kap 6	Eggers & Hainmueller (2009)
14	'Big data' og maskinlæring	Varian (2014), Montgomery & Olivella (2017)	Theocharis et al. (2016)

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*

Frederick Mosteller, Harvard University

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.

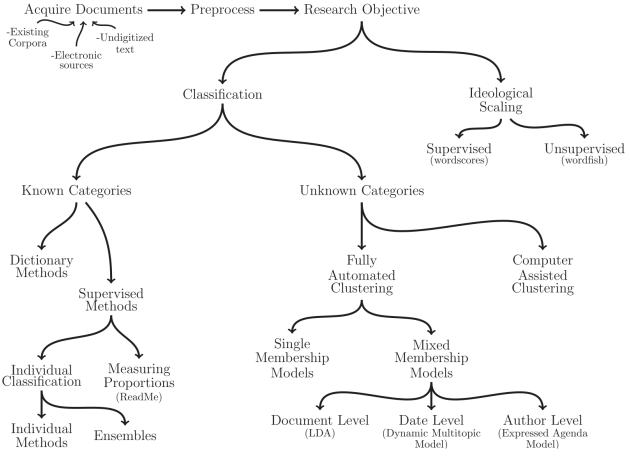
TABLE 2.1. FREQUENCY DISTRIBUTION OF RATE PER THOUSAND WORDS FOR THE 48 HAMILTON AND 50 MADISON PAPERS FOR *by*, *from*, AND *to*. THE UPPER LIMIT OF A CLASS INTERVAL IS NOT INCLUDED IN THE CLASS

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						

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)
$\rightarrow$ denne sondring vender tilbage i sidste holdtime om maskinlæring!

- 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:

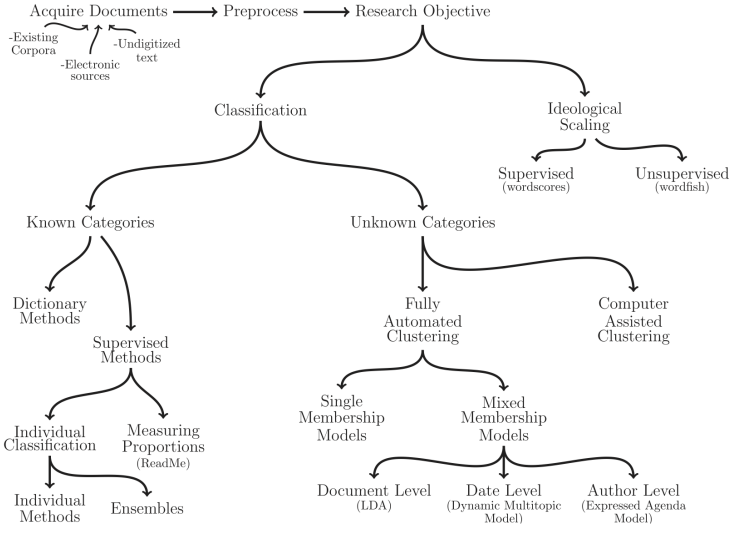
1. import af tekster som et *korpus*
2. pre-processering:
   - fjern tal, specialtegn
   - fjern 'stopwords'
   - stemming mhp. dimensionalitetsreduktion
   - fjern meget sjældne el. hyppige ord
3. konvertering til *document-term/document-feature* matrice
4. analyse

Fra min egen forskning: document-feature matrice med ≈113k folketingstaler

```
> spdfm
Document-feature matrix of: 113,104 documents, 192,155 features (99.9% sparse).
> spdfm[1:10,1:10]
Document-feature matrix of: 10 documents, 10 features (39% sparse).
10 x 10 sparse Matrix of class "dfm"
                                                                features
docs                                                             : regeringens forslag om , at alle skal betale 1
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-10.txt  2           1       1  4 18 11     3     2       2 1
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-100.txt 2           0       0  1 21  7     0     3       0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-101.txt 2           0       0  1 22  5     0     0       0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-102.txt 1           0       0  3 16  6     1     1       0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-103.txt 2           0       1  2 15  7     0     3       0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-104.txt 3           1       0  2 43 18     0     6       0 1
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-105.txt 3           0       0  1  9  7     0     1       0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-106.txt 1           1       1  4 21 12     0     2       0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-107.txt 3           0       0  2 25 14     0     1       0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-108.txt 1           1       1  2 19  6     0     0       0 0
> length(spdfm)
[1] 21733499120
```

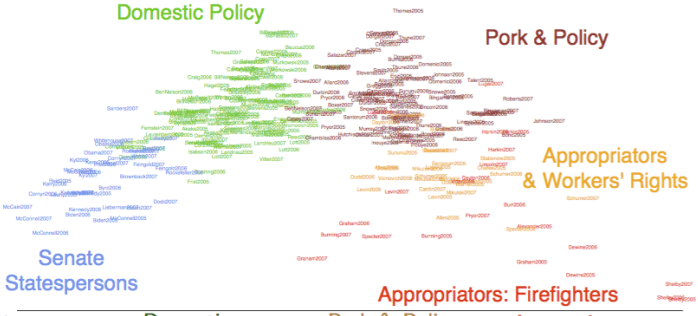
- klassisk pakke til text as data: `tm`
- nyere, enklere alternativ: `quanteda` af Ken Benoit et al.
- fremgangsmåde m. `quanteda`:
  1. import m. `readtext()` i standalone-pakken `readtext`
  2. definition som korpus m. `corpus()`
  3. preprocessering+konvertering m. `dfm()`
  4. analyse, fx. m. `textmodel_*()`

$\rightarrow$ 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
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)

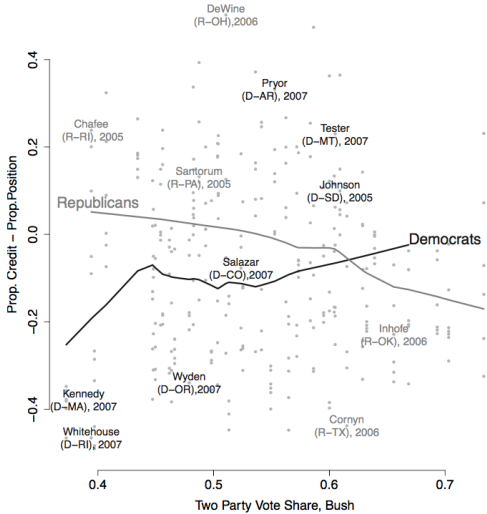
- hvad handler teksterne om?
- $\rightsquigarrow$ hvilke latente kategorier (emner) udspringer teksterne af?
- typisk anvendt approach: emnemodeller (topic models)
- her: *tf-idf* $\rightarrow$ ret primitiv, men letforståelig

Grimmer (2013): Analyse af 64k pressemeddelelser



Senate Position Taker	Domestic Policy	Pork & Policy	Appropriators
- Iraq War	- Environment	- WRDA grants	- Fire Grants
- Intelligence	- Gas prices	- Farming	- Airport Grants
- Intl. Relations	- DHS	- Health Care	- University Money
- Budget	- Consumer Safety	- Education Policy	- 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

$\rightarrow$ hvad er tf-idf for 'velfærd' hos Enhedslisten?

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

Eksempler på dictionaries

- General Inquirer Database (`http://www.wjh.harvard.edu/~inquirer/`)
  - Stone, P.J., Dumphy, D.C., and Ogilvie, D.M. (1966) *The General Inquirer: A Computer Approach to Content Analysis*
  - { Positiv, Negativ }
  - 3627 negative ord positive ord
  - 'workhorse'-ordbog anvendt i mange papers
- Linguistic Inquiry Word Count (LIWC)
  - Tilblivelsesproces:
    - "We drew on common emotion rating scales...Roget's Thesaurus...standard English dictionaries. [then] brain-storming sessions among 3-6 judges were held"⇝ flere ord i samme kategori
  - 2300 ord i 70 kategorier
  - pris: ca. 100USD
- Harvard-IV-4
- Affective Norms for English Words (ANEW)
- AFINN (inkl. dansk ordbog!)

Dictionary-mål for tekster

- Vektor af ordantal i hvert dokument: $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iK})$, $(i = 1, \ldots, N)$
- Vægte til hvert ord $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$
  - $\theta_k \in \{0, 1\}$
  - $\theta_k \in \{-1, 0, 1\}$
  - $\theta_k \in \{-2, -1, 0, 1, 2\}$
  - $\theta_k \in \Re$

For hvert dokument i, udregn scoren

$$Y_i = \frac{\sum_{k=1}^{K} \theta_k X_{ik}}{\sum_{k=1}^{K} X_k} \tag{1}$$

Kontinuert $Y_i \rightsquigarrow$ Klassifikation

$Y_i > 0 \Rightarrow$ Positiv
$Y_i < 0 \Rightarrow$ Negativ
$Y_i \approx 0$ Udefineret

Measuring Happiness

Dodds and Danforth (2009): bruger en dictionary-metode til at måle 'lykke' (a.k.a. 'sentiment analysis')

- Affective Norms for English Words (ANEW)
  - On a scale of 1-9 how happy does this word make you?
    Happy : triumphant (8.82)/paradise (8.72)/ love (8.72)
    Neutral: street (5.22)/ paper (5.20)/ engine (5.20)
    Unhappy : cancer (1.5)/funeral (1.39)/ rape (1.25) /suicide (1.25)
- Lykke for tekst i (med ord j med lykke θ_j og hyppighed X_{ij})

$$\text{Lykke}_i = \frac{\sum_{k=1}^{K} \theta_k X_{ik}}{\sum_{k=1}^{K} X_{ik}}$$

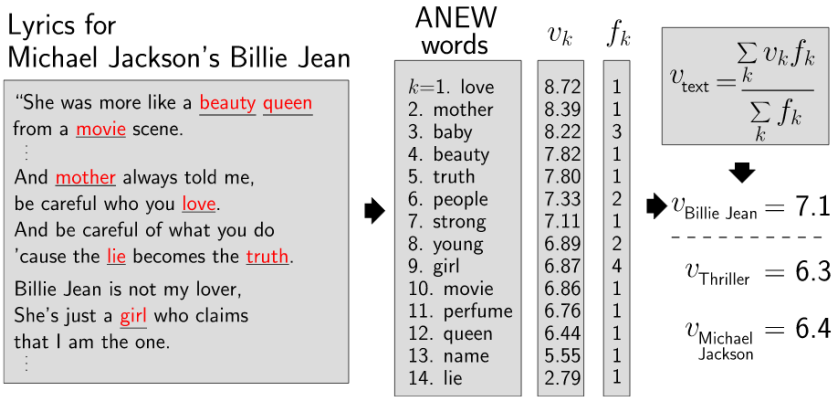


Fig. 2 A schematic example of our method for measuring the average psychological valence of a text, in this case the lyrics of Michael Jackson's Billie Jean. Average valences for the song Billie Jean, the album Thriller, and all of Jackson's lyrics are given at right

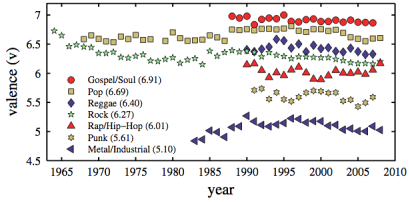


Fig. 6 Valence time series for song titles broken down by representative genres. For each genre, we have omitted years in which less than 1000 ANEW words appear

Table 3 Average valence scores for the top and bottom 10 artists for which we have the lyrics to at least 50 songs and at least 1000 samples of (nonunique) words from the ANEW study word list

Rank	Top artists	Valence	Bottom artists	Valence
1	All 4 One	7.15	Slayer	4.80
2	Luther Vandross	7.12	Misfits	4.88
3	S Club 7	7.05	Staind	4.93
4	K Ci & JoJo	7.04	Slipknot	4.98
5	Perry Como	7.04	Darkthrone	4.98
6	Diana Ross & the Supremes	7.03	Death	5.02
7	Buddy Holly	7.02	Black Label Society	5.05
8	Faith Evans	7.01	Pig	5.08
9	The Beach Boys	7.01	Voivod	5.14
10	Jon B	6.98	Fear Factory	5.15

Dictionary-eksempel: Daisys taler

Problemer med dictionary-metoder

Dictionary-metoder er kontekstinvariante

- optimering $\rightsquigarrow$ modellen tilpasser sig konteksten
- i dictionary-metoder, ingen optimering $\rightsquigarrow$ samme ordvægte uanset kontekst
- $\rightsquigarrow$ modellens performance er usikker

Bare fordi ord er klassificeret som 'positive' eller 'negative' er de ikke nødvendigvis valide mål i sammenhængen $\rightarrow$ husk at validere målet!

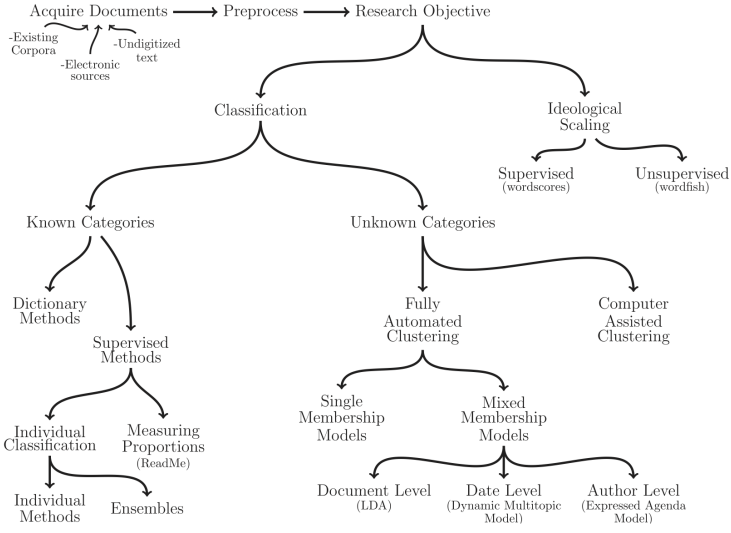
Eks.: anvendelse på dictionary-anvendelse i nye domæner Revisionsforskning: mål af tone i årsrapporter

- omfattende tekstlig sammenfatning af virksomhedens performance
- årsrapportens tone er væsentlig ($)

Tidligere state of the art: Harvard-IV-4 Dictionary
Loughran and McDonald (2011): finansielle dokumenters vokabular er anderledes, præget af polysemi

- Negative ord i Harvard-IV-4, ej negative i revision:
  `tax, cost, capital, board, liability, foreign, cancer, crude (oil), tire`
- 73 pct. af Harvard-IV-4's negative ord er i denne gruppe(!)
- Ej negative i Harvard-IV-4, negative i revision:
  `felony, litigation, restated, misstatement, unanticipated`

$\rightsquigarrow$ Kontekst er afgørende


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:

$$\hat{\theta}_d = \frac{1}{W} \sum_{w=1}^{W} \hat{\pi}_w \tag{2}$$

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

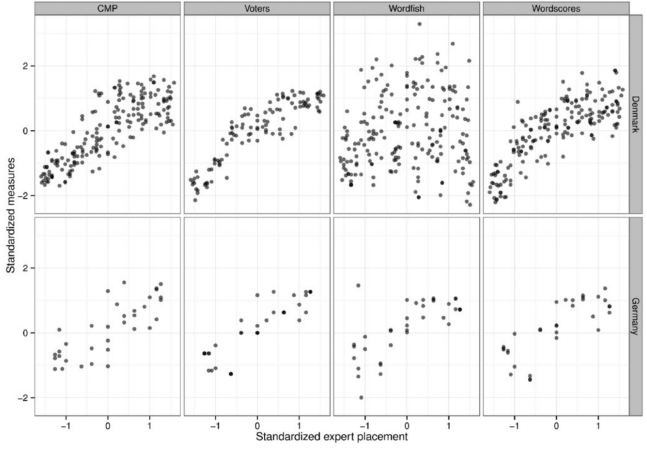
$$\hat{\pi}_w = \sum_{r=1}^{R} \theta_r \hat{P}(d_r|w) \tag{3}$$

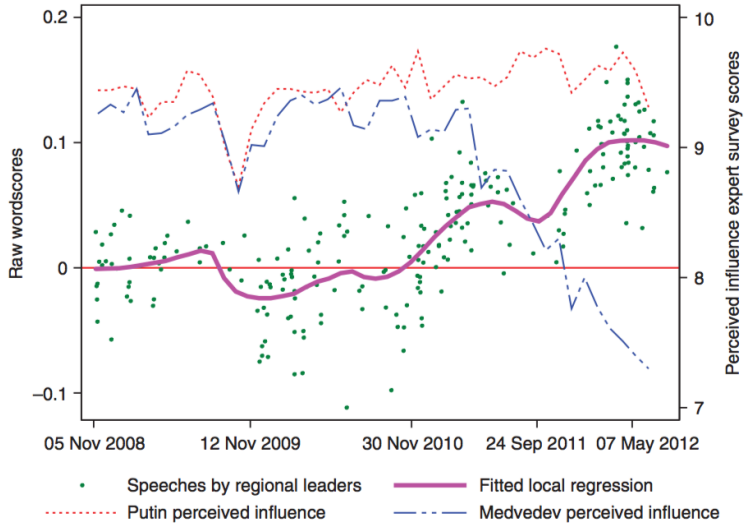
hvor pr. Bayes' teorem:

$$\hat{P}(d_r|w) = \frac{\hat{P}(w|d_i)}{\sum_{r=1}^{R} \hat{P}(w|d_r)} \tag{4}$$

$\rightarrow$ 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: OLS

- husk: ikke næste torsdag, men onsdag d. 10. 15-17 lok. 7.0.18
- læs MM kap. 2 om regression
- læs Mutz (case-tekst til både OLS og panel)
- øvelse til næste gang:
  1. definér din egen ordbog
  2. brug den til at analysere Daisys nytårstaler ligesom i eksemplet
  3. vurder om variationen har face-validitet

Tak for i dag!