14: 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

12. december 2016

1. Formalia
2. Opsamling fra sidst
3. Intro til text as data
4. Klassifikation
   - tf-idf
5. Skalering
   - Wordscores
6. Kig fremad

Uge	Dato	Tema	Litteratur	Case
1	5/9	Introduktion til R	Imai kap 1	
2	12/9	Regression I: OLS	GH kap 3, MM kap 2	Gilens & Page (2014
3	26/9	Regression II: Paneldata	GH kap 11	Larsen et al. (2016)
4	29/9	Regression III: Multileveldata, interaktioner	GH kap 12	Berkman & Plutzer
5	3/10	Introduktion til kausal inferens	Hariri (2012), Samii (2016)	
6	10/10	Matching	Justesen & Klemmensen (2014)	Ladd & Lenz (2009)
	17/10	*Efterårsferie*		

Uge	Dato	Tema	Litteratur	Case
	17/10	*Efterårsferie*		
7	24/10	Eksperimenter I	MM kap 1, GG kap 1+2	Gerber et al. (2008)
8	31/10	Eksperimenter II	GG kap 3+4+5	Gerber & Green (200
9	14/11	Instrumentvariable	MM kap 3	Arunachalam & Wats
10	14/11	Regressionsdiskontinuitetsdesigns	MM kap 4	Eggers & Hainmuelle
11	21/11	Difference-in-difference designs	MM kap 5	Enos (2016)
12	28/11	'Big data' og maskinlæring	Harford, Grimmer, Varian, Athey/Imbens	
13	5/12	Webscraping	MRMN kap 9+14	
14	12/12	Tekst som data	Grimmer & Stewart (2013), Imai kap 5	

- screen scraping ctr. API'er
- screen scraping m. `rvest`
- etik i scraping
- generelt om API'er
- brug af Twitters REST API m. `twitteR`
- case: skalering af danske Twitter-brugere

Spørgsmål?

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

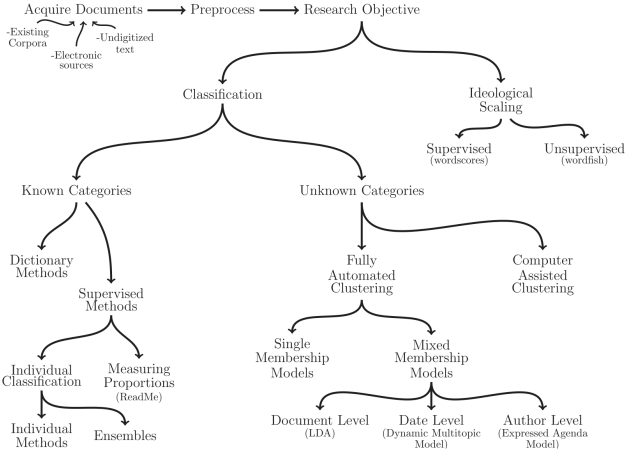
A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers

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

Overordnet sondring:

- klassifikation → hvad handler teksterne om
- skalering → hvordan er teksterne fordelt på en skala

+ i begge tilgange en velkendt sondring: supervised ctr. unsupervised



- 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, lille 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 hos lmai:

```
## 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
```

- typisk pakke til text as data: `tm`
- nyere, enklere alternativ: `quanteda` af Ken Benoit et al.
- fremgangsmåde m. `quanteda`:
  1. import m. `corpus()`, evt. mappehenvisning m. `textfile()`
  2. preprocessering+konvertering m. `dfm()`
  3. analyse, fx. m. `textmodel()`

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?

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

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 $\hat{\pi}_w$:

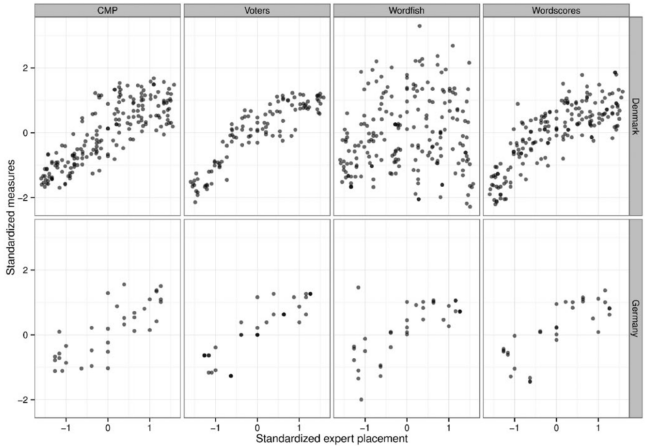
$$\hat{\pi}_w = \sum_{r=1}^{R} \theta_r \hat{P}(d_r|w) \tag{2}$$

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

For partiprogrammer reproducerer Wordscores positioner, cf. Hjorth et al. (2015):



Tak for denne gang!