

13: Webscraping

Videregående kvantitative metoder i studiet af politisk adfærd

Frederik Hjorth

fh@ifs.ku.dk

fghjorth.github.io

@fghjorth

Institut for Statskundskab

Københavns Universitet

5. december 2016

- 1 Formalia
- 2 Opsamling fra sidst
- 3 Screen scraping
- 4 Etik i scraping
- 5 Web API'er
- 6 Case: skalering af danske twitter-brugere
- 7 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 & Hainmueller
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	Web scraping	MRMN kap 9+14	
14	12/12	Tekst som data	Grimmer & Stewart (2013), Imai kap 5	

frivillig workshop i dag kl. 13-15 i Digital Social Science Lab

- definitioner af big data/ML
- hype om big data
- skepsis om big data
- ML: regressionstræer
- ML: LASSO regression
- implementering (hængeparti: LASSO regression m. `glmnet`)

Spørgsmål?

Nyttig pakke til screen scraping: `rvest`

- Typisk fremgangsmåde:
 - ① indlæs html med `read_html()`
 - ② find relevante elementer med `html_nodes()`
 - ③ konverter elementerne til data frames med `html_table()`

Eksempel på scraping m. rvest: den danske kongerække



Gorm den gamle d. n. 940
Gorm the Old d. n. 940



Harald Blåtand o. 940-o. 987
Harald Bluetooth
c. 940-c. 987



Svend Tvedeg o. 986-1014
Svend Forkbeard
c. 986-1014



Knud den Store 1018-1035
Canute the Great



Svend Estriden
1047-1074



Erik Ejend 1095-1103
Erik the Good



Niels den Gamle 1104-1134
Niels the Old



Valdemar den Store 1157-1182
Valdemar the Great



Knud VI 1182-1202
Canute VI



Valdemar Sejor 1202-1241
Valdemar the Victorious



Erik Plovskeping 1241-1250
Erik Ploughshare

Trick: brug 'Inspect...' til at finde HTML/CSS-tegn i kildekoden

Der er for så vidt muligt anvendt samtidige portrætter, men de fleste portrætter i afsnittet er ikke samtidige (normalt fra 1600-tallet) og kan ikke forventes at have nogen lighed med regenten.

Navn	Billede	Født	Tiltrådte	Fratrådte/Død
Gorm den Gamle		Slutningen af 800-tallet evt. mellem 908 og 918. ^[5] Søn af Hardeknud	Ukendt (Nævnes første gang i 936)	Antagelig død 958 mellem 40 og 50 år gammel
Harald Blåtand		Ca. 935 Søn af Gorm den Gamle og Thyra Dannebød	Konge 958 ved Gorm den Gamles død	Død 985, 986 eller 987 ^[6] 50-52 år gammel
Svend Tveskæg		Ca. 960 Søn af Harald Blåtand og Gunhild	Konge 986 eller 987 ^[8] ved Harald	Død 3. februar 1014 i Lincoln i England

Inspect... window showing the HTML structure and CSS styles for the table.

Elements

```

<p>Der er flere muligheder for, hvem Knud den 1. var:</p>
<p>...</p>
<h2>...</h2>
<p>Kongerækken gælder disse områder:</p>
<ul>...</ul>
<h3>...</h3>
<div class="detail">...</div>
<p>...</p>
<p>...</p>
<p>...</p>

```

Styles

```

table.wikitable {
  text-align: center;
  width: 99%;
}

```

Properties

margin: 14px; border: 1px solid #f9f9f9; padding: 10px; width: 649px; height: 611px;

Spørgsmål?

Cautionary tale I: Aaron Schwartz vs. JSTOR



Cautionary tale II: OKCupid data dump

Researchers just released profile data on 70,000 OkCupid users without permission

Updated by Brian Resnick | @B_resnick | brian@vox.com | May 12, 2016, 6:00pm EDT

→ baggrund hos Vox.com


offentligt tilgængeligt → åbne data

Emil OW Kirkegaard @KirkegaardEmil · May 8
The OKCupid paper has now been submitted. This means that the dataset is now public! Enjoy! :3 [openpsych.net/forum/showthre...](https://openpsych.net/forum/showthread.php?p=1000000)


↳ 26 ♡ 36 ...

Ethan Jewett @esjewett · May 11
@KirkegaardEmil This data set is highly re-identifiable. Even includes usernames? Was any work at all done to anonymize it?

↳ 3 ♡ 9 ...

Emil OW Kirkegaard @KirkegaardEmil  [Follow](#)

@esjewett No. Data is already public.

LIKE 1 

1:30 PM - 11 May 2016

↳ ♡ ...

 Reply to @KirkegaardEmil @esjewett


Ethan Jewett @esjewett · May 11
@KirkegaardEmil Differing degrees of "public". Also different ethical guidelines. IMO, you should speak with a research ethicist/IRB ASAP.

↳ 2 ♡ 21 ...


SAM @samuelmehr · May 11 

@BrianNosek @KirkegaardEmil super cool. interesting that the data are scraped and not provided by @okcupid, are they OK with this?

↳ 1 ♡ ...

Emil OW Kirkegaard @KirkegaardEmil  [Follow](#)

@samuelmehr @BrianNosek Don't know, don't ask. :)

RETWEET 1 

11:08 AM - 11 May 2016

↳ ♡ ...

Spørgsmål?

API: Application Programming Interface

»tools which enable programmers to connect their software with 'something else'« (p. 259)

- essentiel byggesten i 'Web 2.0.'
- vigtigt her: værktøj til at rekvirere data

Autorisering

- de færreste API'er er åbne
- typisk autoriseringsstandard: OAuth
- bruges fx. til at autorisere databrug i tredjepartsapps
- kræver *client credentials* for at tilgå data fra API
- typisk procedure:
 - ① registrér app på API'ens side
 - ② skaf *consumer key*+*secret*
 - ③ autorisér vha. *key*+*secret* i R

REST: REpresentational State TTransfer

- API-standard udviklet til at hente statiske data
- bruger standard HTTP-metoder såsom GET, POST
- API'er der opfylder REST-standarder kaldes RESTful
- mange nyttige R-pakker er wrapper-funktioner for API'er

Twitter's REST API

- kan tilgås med pakken `twitterR`
- consumer key/secret kan tilgås på `apps.twitter.com`
- rate limits: afhængig af type data 15 el. 180 requests pr. 15 min.
- rate limit status kan tjekkes m. `getCurRateLimitInfo()`

Muligheder og begrænsninger i REST API'en

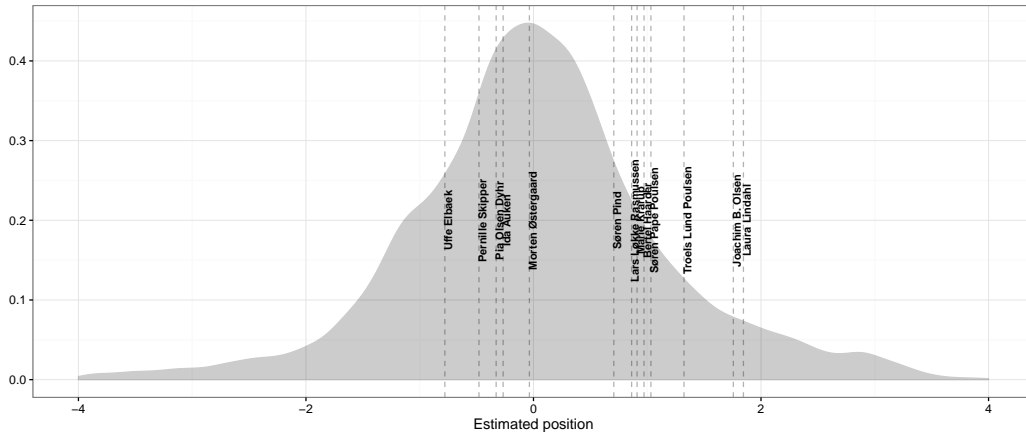
- kan tilgå information om relationer ml. konti
- kan tilgå alle tweets for en given bruger
- kan kun returnere emneordssøgninger 7 dage bagud
- alternativ: gem løbende tweets fra streaming API'er vha. `streamR`

Mål: estimér politiske positioner hos twitter-brugere (\rightarrow sml. m. online adfærd)

Fremgangsmåde:

- ① find liste med danske folketingsmedlemmer på Twitter ($N = 154$)
- ② saml vha. REST API liste over alle unikke followers ($N = 164,041$)
- ③ konstruér *adjacency matrix* med relationer ml. alle MF'ere og followers
- ④ reducér adjacency matrix til followers af 10+ MF'ere ($N = 20,091$)
- ⑤ estimér positioner for MF'ere og followers vha. multidimensionel skalering

Est. positioner af udvalgte MP'er + 20k followers



Spørgsmål?

Næste gang:

- tekst som data
- Grimmer & Stewart: god oversigtstekst
- Imai: fokus på 5.1 (resten interesselæsning)
- Benoit & Nulty introducerer R-pakken quanteda (ctr. tm)
- Hjorth et al. (2015) sammenligner skaleringsmetoder, kan læses kursorisk

Tak for i dag!