STV2022 – Store tekstdata

Solveig Bjørkholt og Martin Søyland

2022-06-22

Contents

Chapter 1

Introduksjon

Velkommen til STV2022 – Store teksdata!

Dette er en arbeidsbok som går gjennom de forskjellige delene i kurset, med tilhørende R-kode. Meningen med arbeidsboken, er at den kan brukes som forslag til implementering av metoder i semesteroppgaven. Merk likevel at dette ikke er en fasit!

Chapter 2

Om kurset

I kurset skal vi bli kjent med analyseprosessen av store tekstdata: Hvordan samler man effektivt og redelig store mengder politiske tekster? Hva må til for å gjøre slike tekster klare for analyse? Og hvordan kan vi analysere tekstene?

Politikere og politiske partier produserer store mengder tekst hver dag. Om det er gjennom debatter, taler på Stortinget, lovforslag fra regjeringen, høringer, offentlige utredninger med mer, er digitaliserte politiske tekster i det offentlige blitt mer tilgjengelig de siste tiårene. Dette har åpnet et mulighetsrom for tekstanalyse som ikke var mulig/veldig vanskelig og tidkrevende før.

Det kan ofte være vanskelig å finne mønster som kan svare på spørsmål og teorier vi har i statsvitenskap i disse store tekstsamlingene. Derfor kan vi se til metoder innenfor maskinlæring for å analysere store samlinger av tekst systematisk. Samtidig er ikke alltid digitaliserte politiske tekster tilrettelagt for å analysers direkte. I disse tilfellene er god strukturering av rådata viktig.

Gjennom å delta i dette kurset vil du lære å søke i store mengder dokumenter, oppsummere disse på meningsfulle måter og indentifisere riktige analysemetoder for å teste statsvitenskaplige teorier med store tekstdata. Kurset vil dekke samling av store volum tekst fra offentlige kilder, strukturering og klargjøring av tekst for analyse og kvantitative tekstanalysemetoder.

Chapter 3

Oppbygging av arbeidsboken

Under vil vi gå gjennom undervisningsopplegget, som arbeidsboken er lagt opp etter. Delene av boken er strukturert som følgende:

1. Anskaffelse av tekst
2. Laste inn eksisterende tekstkilder
3. Forbehandling av tekst (preprosessering)
4. Veiledet læring (supervised)
5. Ikke-veiledet læring (unsupervised)
6. Ordbøker
7. Tekstsatistikk
8. Sentiment
9. Temamodellering
10. Latente posisjoner i tekst

Chapter 4

Anbefalte forberedelser

Siden kurset krever noe forkunnskap om R og generell metodisk kompetanse, anbefaler vi å se over følgende materiale før kurset starter:

- Arbeidsbøker for R ved UiO
- R materiale for STV1020

Chapter 5

Undervisning

Undervisningen i STV2022 består av 10 forelesninger og 5 seminarer. Vi vil bruke forelesningene til å oppsummere hovedkonseptene i hver ukes tema, både metodisk og anvendt. Seminarene vil ha hovedfokus på teknisk gjennomføring av tekstanalyse i R. Hvert seminar vil være delt i to med én del der seminarleder går gjennom ekstempler på kodeimplementering og én del der studentene kan jobbe med semesteroppgaven. Det er også verdt å merke seg at mange av implementeringene i kurset krever en del prøving og feiling.

Merk at det etter hvert seminar skal leveres inn et utkast av oppgaven for temaet man har gått gjennom i seminaret. Disse delene må bestås for å få vurdert semesteroppgave.

5.1 Forelesninger

De ti forelesningene har følgende timeplan (høsten 2022):

Dato	Tid	Aktivitet	Sted	Foreleser	Ressurser/pensum
ti. 23. aug.	10:15–12:00	Introduksjon	ES, Aud. 5	S. Bjørkholt og M. Søyland	Grimmer et al. [2022] kap. 1-2 og 22, Lucas et al. [2015], Silge and Robinson [2017] kap. 1, Pang et al. [2008] kap. 1
ti. 30. aug.	10:15–12:00	Anskaffelse og innlasting av tekst	ES, Aud. 5	M. Søyland	Grimmer et al. [2022] kap. 3-4, Cooksey [2014] kap. 1, Wickham [2020], Høyland and Søyland [2019]

Dato	Tid	Aktivitet	Sted	Foreleser	Ressurser/pensum
ti. 6. sep.	10:15–12:00	Forbehandling av tekst 1	ES, Aud. 5	M. Søyland	Grimmer et al. [2022] kap. 5, Silge and Robinson [2017] kap. 3, Jørgensen et al. [2019], Barnes et al. [2019], Benoit and Matsuo [2020]
ti. 13. sep.	10:15–12:00	Forbehandling av tekst 2	ES, Aud. 5	S. Bjørkholt	Grimmer et al. [2022] kap. 9, Silge and Robinson [2017] kap. 4, Denny and Spirling [2018]

Dato	Tid	Aktivitet	Sted	Foreleser	Ressurser/pensum
ti. 20. sep.	10:15–12:00	Bruke API – Case: Stortinget	ES, Aud. 5	M. Søyland	Stortinget [2022], Søyland [2022], Finseraas et al. [2021]
ti. 11. okt.	10:15–12:00	Veiledet og ikke-veiledet læring	ES, Aud. 5	S. Bjørkholt	Grimmer et al. [2022] kap. 10 og 17, D'Orazio et al. [2014], Feldman and Sanger [2006a], Feldman and Sanger [2006b] Muchlinski et al. [2016]

Dato	Tid	Aktivitet	Sted	Foreleser	Ressurser/pensum
ti. 18. okt.	10:15–12:00	Ordbøker, tekstlikhet og sentiment	ES, Aud. 5	S. Bjørkholt	Grimmer et al. [2022] kap. 7 og 16, Silge and Robinson [2017] kap. 2, Pang et al. [2008] kap. 3-4, Liu [2015], Liu2015a
ti. 25. okt.	10:15–12:00	Temamodellering	ES, Aud. 5	M. Søyland	Grimmer et al. [2022] kap. 13, Blei [2012], Silge and Robinson [2017] kap. 6, Roberts et al. [2014]

Dato	Tid	Aktivitet	Sted	Foreleser	Ressurser/pensum
ti. 1. nov.	10:15–12:00	Estimere latent posisjon fra tekst	ES, Aud. 5	S. Bjørkholt	Laver et al. [2003], Slapin and Proksch [2008], Lowe [2017], Lauderdale and Herzog [2016], Peterson and Spirling [2018]
ti. 15. nov.	10:15–12:00	Oppsummering	ES, Aud. 5	S. Bjørkholt og M. Søyland	Grimmer et al. [2022] kap 28, Wilkerson and Casas [2017]

5.2 Seminarer

Uke	Aktivitet
36	Seminar 1: Anskaffe tekst og lage dtm i R
38	Seminar 2: Preprosessering av tekstdata i R
42	Seminar 3: Veiledet og ikke-veiledet læring i R
44	Seminar 4: Modelleringsmetoder i R
46	Seminar 5: Fra tekst til funn, Q&A og oppgavehjelp

Seminarledere:

- Eli Sofie Baltzersen elibal@student.sv.uio.no
- Eric Gabo Ekeberg Nilsen e.g.e.nilsen@stv.uio.no

5.3 Pensum

Som med alle andre fag, er det sterkt anbefalt at man ser over pensum før forelesning og seminar. Likevel kan pensum i kurset til tider være noe teknisk og uhåndterbart. Det er ikke forventet å *pugge* formler eller fult ut forstå de matematiske beregninger bak de forskjellige modelleringsmetodene (selv om det åpenbart kan gjøre stoffet lettere å forstå). Hovedfokuset vårt vil være på å forstå hvilke operasjoner man må gjøre for å gå fra tekst til funn, hvilke antagelser man gjør i prosessen og klare å velge de riktige modellene for spørsmålet man vil ha svar på.

Grunnboken i pensum er Grimmer et al. [2022]. Vi vil lene oss mye på denne over alle temaene vi gjennomgår. For R har vi valgt å gjøre materialet så standardisert som mulig ved å bruke `tidyverse` så langt det lar seg gjøre. Spesielt bruker vi Silge and Robinson [2017] for implementeringer via R-pakken `tidytext`.

Vi har også lagt inn noen bidrag som anvender metodene vi går gjennom i løpet av kurset, som Peterson and Spirling [2018], Lauderdale and Herzog [2016], Høyland and Søyland [2019], Finseraas et al. [2021], for å synliggjøre nytten av metodene i anvendt forskning.

Chapter 6

Laste inn tekstdata

Om å laste tekst!

Chapter 7

Anskaffelse av tekst

7.1 .html-skraping

Dette kan være lite gøy

7.2 .xml-skraping

Dette er enklere enn html

7.3 .json-skraping

Dette liker jeg ikke. Veldig rar datastrukturering

7.4 Curl

Ja, må vi egentlig snakke om curl

7.5 APIer

Kanskje bruke Stortinget som eksempel.

Chapter 8

Preprosessering

Preprosessering er ganske viktig.

8.1 Sekk med ord

Alle tekster er unike!

Ta for eksempel spor 6 på No.4-albumet vi allerede har jobbet med – *Regndans i skinnjakke*. Hvis vi skal følge en vanlig antagelse i kvantitativ tekstanalyse – "sekk med ord" eller *bag of words* – skal vi kunne forstå innholdet i en tekst hvis vi deler opp teksten i segmenter, putter det i en pose, rister posen og tømmer det på et bord. Da vil denne sangen for eksempel se slik ut:

```
regndans <- readLines("./data/regndans.txt")

bow <- regndans %>%
    str_split("\\s") %>%
    unlist()

set.seed(984301)

cat(bow[sample(1:length(bow))])
```

```
## begynner kaffe i på Ta backflip Prøver rustfarva, når Gresstrå Drikke skinnjakke er I på I TV-
```

De fleste (som ikke kan sangen fra før) vil ha vanskelig å forstå hva den egentlig handler om bare ved å se på dette. Vi kan identifisere meningsbærende ord som "Oslofjorden", "Grille", "trampoline", "dragepust", med mer. Likevel er det vanskelig å skjønne hva låtskriveren egentlig vil formidle med denne teksten. Det er dette som gjør "sekk med ord"-antagelsen veldig streng. Språk er veldig komplekst og ordene i en tekst kan endre mening drastisk bare ved å se på

en liten del av konteksten de dukker opp i. Om vi bare ser på linjen som inneholder orded "dragepust", innser vi fort at konteksten rundt ordet gir oss et veldig tydelig bilde av hva låtskriveren mener med akkurat den linjen:

```
regndans[which(str_detect(regndans, "dragepust"))]
```

```
## [1] "Våkne opp m d dragepust"
```

Likevel gir det oss ikke et godt bilde på hva teksten handler om i sin helhet. Det får vi bare sett ved å se på hele teksten:

```
## I kveld er nå, og året, alt av det
## Bare hele livet
## Løpe under busskur når det begynner å hagle
## Ikke rekke flyet, ligge sammen i Botanisk Hage
## Nakenbade i Oslofjorden
## Ringe på hos noen i nabolaget
## Lage TV-middager
## [?]
## Hente i barnehager, altså
## Regndanse i skinnjakke
## Ta T-banen til Hasle
## Drikke hundre glass med øl, ass
## Tusen kopp r kaffe
## Grille bratwürst på [?]
## Våkne opp m d dragepust
## Se varmluftsballonger
## Bjørkeblader i august blir gule
## Også rustfarva, og løsne og fly avgårde
## Gresstrå på høsten, ass
## Hårfestet begynner å gråne
## Gå hjem og går forbi
## En gutt tar backflip på en trampoline
## Se endorfinene krystalliserer seg i smilehulla dine
## Prøver jeg å si
## I kveld er hele livet
```

Nå teksten gir mening! Tolkninger kan selvfølgelig variere fra individ til individ og den "riktige" tolkningen, er det bare forfatteren som vet hva er. Personlig tolker jeg denne teksten som et utløp for frustrasjon under corona-pandemien, og prospektene ved livet når samfunnet gjenåpnes, fordi jeg hørte den for første gang under nedstengningen.

Hovedpoenget med å vise dette er at *sekk med ord*-antagelsen er veldig streng og ofte veldig urealistisk. Tekster (og språk generelt) er ekstremt komplekst. Det kan variere mellom geografiske områder (nasjoner, dialekter, osv), aldersgrupper, arenaer (talestol, dialog, monolog, osv), og individuell stil. Oppi alt dette skal vi prøve å finne mønster som sier noe om likhet/ulikhet mellom tekster. Heldigvis

har vi flere verktøy som kan hjelpe oss i å lette litt på *sekk med ord*-antagelsen. Men antagelsen vil likevel alltid være der, i en eller annen form. La oss se litt på hvilke teknikker vi kan bruke for å gjøre modellering av tekst noe mer omgripelig¸ men aller først skal vi se litt på hvilke trekk som muligens ikke gir oss så mye informasjon om det vi er ute etter, eller støy, som vi ofte vil fjerne.

8.2 Fjerne trekk?

8.2.1 Punktsetting

8.2.2 Stoppord

8.3 Rotform av ord

8.3.1 Stemming

8.3.2 Lemmatisering

8.4 ngrams

8.5 Taledeler (parts of speech)

Chapter 9

Veildedet læring

Chapter 10

Ikke-veiledet læring

Chapter 11

Ordbøker

Chapter 12

Tekststatistikk

12.1 Likhet

12.2 Avstand

12.3 Lesbarhet

12.4 Uttrykk

Chapter 13

Sentiment

13.1 NorSentLex

Chapter 14

Temamodellering

Chapter 15

Latente posisjoner

Chapter 16

Oppsummering

Bibliography

Jeremy Barnes, Samia Touileb, Lilja Øvrelid, and Erik Velldal. Lexicon information in neural sentiment analysis: a multi-task learning approach. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 175–186, Turku, Finland, September–October 2019. Linköping University Electronic Press. URL https://aclanthology.org/W19-6119.

Kenneth Benoit and Akitaka Matsuo. *spacyr: Wrapper to the 'spaCy' 'NLP' Library*, 2020. URL https://CRAN.R-project.org/package=spacyr. R package version 1.2.1.

David M Blei. Probabilistic Topic Models. *Communications of the ACM*, 55(4): 77 – 84, 2012.

Brian Cooksey. An introduction to APIs. *Zapier, Inc.*, 2014. URL https://cdn.zapier.com/storage/learn_ebooks/e06a35cfcf092ec6dd22670383d9fd12.pdf.

Matthew J. Denny and Arthur Spirling. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, 26(2):168–189, 2018. URL https://doi.org/10.1017/pan.2017.44.

Vito D'Orazio, Steven T. Landis, Glenn Palmer, and Philip Schrodt. Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political Analysis*, 22(2):224 – 242, 2014. doi: https://doi.org/10.1093/pan/mpt030.

Ronen Feldman and James Sanger. *Categorization*, pages 64–81. Cambridge University Press, 2006a. doi: 10.1017/CBO9780511546914.005.

Ronen Feldman and James Sanger. *Clustering*, pages 82–93. Cambridge University Press, 2006b. doi: 10.1017/CBO9780511546914.006.

Henning Finseraas, Bjørn Høyland, and Martin G. Søyland. Climate politics in hard times: How local economic shocks influence mps attention to climate change. *European Journal of Political Research*, 60(3):738–747, 2021. URL https://ejpr.onlinelibrary.wiley.com/doi/abs/10.1111/1475-6765.12415.

Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. *Text as Data:*

A New Framework for Machine Learning and the Social Sciences. Princeton University Press, 2022.

Bjørn Høyland and Martin Søyland. Electoral reform and parliamentary debates. *Legislative Studies Quarterly*, 44(4):593–615, 2019. doi: 10.1111/lsq.12237.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. Norne: Annotating named entities for norwegian, 2019. URL https://arxiv.org/abs/1911.12146.

Benjamin E. Lauderdale and Alexander Herzog. Measuring political positions from legislative speech. *Political Analysis*, 24(3):374–394, 2016. doi: 10.1093/pan/mpw017.

Michael Laver, Kenneth Benoit, and John Garry. Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, 97(02):311–331, 2003.

Bing Liu. Introduction. In *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, pages 1–15. Cambridge University Press, 2015. ISBN 978-1-139-08478-9. URL https://www.cambridge.org/core/books/sentiment-analysis/3F0F24BE12E66764ACE8F179BCDA42E9.

Will Lowe. Understanding wordscores. *Political Analysis*, 16(4):356–371, 2017. doi: 10.1093/pan/mpn004.

Christopher Lucas, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, 23(2):254–277, 2015. doi: 10.1093/pan/mpu019.

David Muchlinski, David Siroky, Jingrui He, and Matthew Kocher. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1):87–103, 2016. doi: 10.1093/pan/mpv024.

Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135, 2008. URL https://www.cs.cornell.edu/home/llee/omsa/omsa.pdf.

Andrew Peterson and Arthur Spirling. Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems. *Political Analysis*, 26(1), 2018.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082, Mar 2014. ISSN 0092-5853.

Julia Silge and David Robinson. *Text mining with R: A tidy approach*. O'Reilly Media, Inc., 2017. URL https://www.tidytextmining.com/.

Jonathan B. Slapin and Sven-Oliver Proksch. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722, Jul 2008. ISSN 1540-5907.

Stortinget. *Stortingets datatjeneste*, 2022. URL https://data.stortinget.no.

Martin Søyland. *stortingscrape: Scrape and structure raw data from the Norwegian parliament's API*, 2022. URL https://github.com/martigso/stortingscrape.

Hadley Wickham. *httr: Tools for Working with URLs and HTTP*, 2020. URL https://cran.r-project.org/web/packages/httr/vignettes/quickstart.html. R package version 1.4.2.

John Wilkerson and Andreu Casas. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20(1):529–544, 2017. URL https://doi.org/10.1146/annurev-polisci-052615-025542.