STV2022 – Store tekstdata

Solveig Bjørkholt og Martin Søyland

2022-06-01

Contents

Chapter 1

Introduksjon

Velkommen til STV2022 – Store teksdata!

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 er tilrettelagt for å analysers direkte. I disse tilfellene er god strukturering av rådata viktig.

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. Kurset vil dekke samling av store volum tekst fra offentlige kilder, strukturering og klargjøring av tekst for analyse og kvantitative tekstanalysemetoder.

Chapter 2

Undervisning

Under beskrives undervisningen i STV2022.

2.1 Forelesninger

Kurset har 10 forelesninger:

1. Intro! (uke 34)
2. Anskaffelse og innlasting av tekst (uke 35)
3. Forbehandling av tekst 1 (uke 36)
4. Forbehandling av tekst 2 (uke 37)
5. Bruke API (Stortinget) (uke 38)
6. Veiledet versus ikke-veiledet læring (uke 41)
7. Ordbøker, tekstlikhet og sentiment (uke 42)
8. Klassifisering av tekst – temamodellering (uke 43)
9. Estimere latent posisjon fra tekst (uke 44)
10. Oppsummering! (uke 46)

Det er sterkt anbefalt at man ser over pensum før forelesning og seminar.

2.2 Pensum

2.2.1 Generelt

Under er en liste over bidrag som vil bli dekket kontinuerlig gjennom hele kurset:

1. Silge and Robinson [2017]
2. Grimmer et al. [2022]
3. Benoit et al. [2017]
4. Jurafsky and Martin [2021] (anbefalt)
5. Wickham [2016] (anbefalt)

2.2.2 Uke 34. Introduksjon (Solveig og Martin)

Den første uken vil vi fokusere på de genrelle konseptene innenfor kvantitativ tekstanalyse, prosessen med å gå fra rå tekst til slutningsanalyse, potensielle fallgruver, med mer.

1. Grimmer et al. [2022] kap. 1-2 og 22 (36 sider)
2. Lucas et al. [2015] (24 sider)
3. Silge and Robinson [2017] kap. 1 (9 sider)
4. Pang et al. [2008] kap. 1 (10 sider)

2.2.3 Uke 35. Anskaffelse og innlasting av tekst (Martin)

1. Grimmer et al. [2022] kap. 3-4 (14 sider)
2. Cooksey [2014] kap. 1 (4 sider)
3. Wickham [2020] (8 sider)
4. Høyland and Søyland [2019] (22 sider)

2.2.4 Uke 36. Forbehandling av tekst 1 (Martin)

1. Grimmer et al. [2022] kap. 5 (11 sider)
2. Silge and Robinson [2017] kap. 3 (9 sider)
3. Jørgensen et al. [2019] (10 sider)
4. Barnes et al. [2019] (12 sider)
5. Benoit and Matsuo [2020] (11 sider)

Seminar 1: Anskaffe tekst og lage dtm i R

2.2.5 Uke 37. Forbehandling av tekst 2 (Solveig)

1. Grimmer et al. [2022] kap. 9 (7 sider)
2. Silge and Robinson [2017] kap. 4 (15 sider)
3. Denny and Spirling [2018] (21 sider)

2.2.6 Uke 38. Bruke API – Case: Stortinget (Martin)

1. Stortinget [2022] (1 sider)
2. Søyland [2022] (1 sider)
3. Finseraas et al. [2021] (10 sider)

Seminar 2: Preprosessering av tekstdata i R

2.2.7 Uke 39. INGEN UNDERVISNING

2.2.8 Uke 40. INGEN UNDERVISNING

2.2.9 Uke 41. Veiledet læring og ikke-veiledet læring (Solveig)

1. Grimmer et al. [2022] kap. 10 og 17 (24 sider)
2. D'Orazio et al. [2014] (18 sider)
3. Feldman and Sanger [2006a] (17 sider)
4. Feldman and Sanger [2006b] (11 sider)
5. Muchlinski et al. [2016] (16 sider)

2.2.10 Uke 42. Ordbøker, tekstlikhet og sentiment (Solveig)

1. Grimmer et al. [2022] kap. 7 og 16 (12 sider)
2. Silge and Robinson [2017] kap. 2 (11 sider)
3. Pang et al. [2008] kap. 3-4 (26 sider)
4. Liu [2015a] (15 sider)
5. Liu [2015b] (30 sider)

Seminar 3: Sup vs. unsup i R

2.2.11 Uke 43. Klassifisering av tekst – Temamodellering (Martin)

1. Grimmer et al. [2022] kap. 13 og (13 sider)
2. Blei [2012] (8 sider)
3. Silge and Robinson [2017] kap. 6 (14 sider)
4. Roberts et al. [2014] (19 sider)

2.2.12 Uke 44. Estimere latent posisjon fra tekst (Solveig)

1. Laver et al. [2003] (20 sider)
2. Slapin and Proksch [2008] (18 sider)
3. Lowe [2017] (15 sider)
4. Lauderdale and Herzog [2016] (20 sider)
5. Peterson and Spirling [2018] (8 sider)

Seminar 4: Modelleringsmetoder i R

2.2.13 Uke 46. Oppsummering (Solveig og Martin)

1. Grimmer et al. [2022] kap 28 (5 sider)
2. Wilkerson and Casas [2017] (19 sider)

Seminar 5: Fra tekst til funn – Q&A/Oppg.hjelp

2.3 Seminarer

Seminarene vil bli

2.4 Nyttige ressurser

- Arbeidsbøker for R ved UiO
- R materiale for STV1020

Chapter 3

Anskaffelse av tekst

Chapter 4

Laste inn tekst

Chapter 5

Preprosessering

Preprosessering er ganske viktig.

5.1 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:

```
library(stringr)

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

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

set.seed(984301)

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

```
##   [1] "begynner"          "kaffe"             "i"
##   [4] "på"                "Ta"                "backflip"
##   [7] "Prøver"            "rustfarva,"        "når"
##  [10] "Gresstrå"          "Drikke"            "skinnjakke"
##  [13] "er"                "I"                 "på"
##  [16] "I"                 "TV-middager"       "av"
##  [19] "Bare"              "Se"                "med"
##  [22] "krystalliserer"    "m d"               "hele"
```

```
##  [25] "Se"                "Bjørkeblader"      "hele"
##  [28] "i"                 "i"                 "hjem"
##  [31] "i"                 "smilehulla"        "jeg"
##  [34] "livet"             "Tusen"             "varmluftsballonger"
##  [37] "noen"              "dine"              "det"
##  [40] "i"                 "[?]"               "nå,"
##  [43] "opp"               "avgårde"           "bratwürst"
##  [46] "det"               "endorfinene"       "Hårfestet"
##  [49] "Gå"                "Hasle"             "gule"
##  [52] "høsten,"           "ass"               "Oslofjorden"
##  [55] "gutt"              "og"                "barnehager,"
##  [58] "alt"               "og"                "løsne"
##  [61] "busskur"           "å"                 "året,"
##  [64] "[?]"               "Også"              "til"
##  [67] "Regndanse"         "T-banen"           "altså"
##  [70] "hundre"            "livet"             "Hente"
##  [73] "gråne"             "glass"             "blir"
##  [76] "rekke"             "begynner"          "Våkne"
##  [79] "dragepust"         "forbi"             "er"
##  [82] "hagle"             "tar"               "å"
##  [85] "kopp r"            "i"                 "Løpe"
##  [88] "på"                "å"                 "Hage"
##  [91] "Lage"              "si"                "En"
##  [94] "øl,"               "Ikke"              "og"
##  [97] "en"                "ass"               "flyet,"
## [100] "sammen"            "nabolaget"         "trampoline"
## [103] "ligge"             "Ringe"             "og"
## [106] "kveld"             "i"                 "fly"
## [109] "under"             "Nakenbade"         "går"
## [112] "Grille"            "kveld"             "hos"
## [115] "på"                "seg"               "august"
## [118] "Botanisk"
```

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.

5.2 Fjerne trekk?

5.2.1 Punktsetting

5.2.2 Stoppord

5.3 Rotform av ord

5.3.1 Stemming

5.3.2 Lemmatisering

5.4 ngrams

5.5 Taledeler (parts of speech)

Chapter 6

Veildedet læring

Chapter 7

Ikke-veiledet læring

Chapter 8

Ordbøker

Chapter 9

Tekststatistikk

9.1 Likhet

9.2 Avstand

9.3 Lesbarhet

9.4 Uttrykk

Chapter 10

Sentiment

10.1 NorSentLex

Chapter 11

Temamodellering

Chapter 12

Latente posisjoner

Chapter 13

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.

Kenneth Benoit, Kohei Watanabe, Paul Nulty, Adam Obeng, Haiyan Wang, Benjamin Lauderdale, and Will Lowe. *quanteda: Quantitative Analysis of Textual Data*, 2017. URL http://quanteda.io. R package version 0.9.9-65.

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.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Online draft, 2021. URL https://web.stanford.edu/~jurafsky/slp3/.

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, 2015a. ISBN 978-1-139-08478-9. URL https://www.cambridge.org/core/books/sentiment-analysis/3F0F24BE12E66764ACE8F179BCDA42E9.

Bing Liu. The problem of sentiment analysis. In *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, pages 16–46. Cambridge University Press, 2015b. ISBN 978-1-139-08478-9. URL https://www.cambridge.org/core/services/aop-cambridge-core/content/view/A0AFE2C49D72C5914C34DAC763BCD931/9781139084789c2_p16-46_CBO.pdf/problem_of_sentiment_analysis.pdf.

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. *ggplot2: elegant graphics for data analysis*. Springer, 2016. URL https://ggplot2-book.org/.

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.