

STV2XXX

Store tekstdata:

Samling, strukturering og analyse av politiske tekster

Martin Søyland* og Solveig Bjørkholt†

November 24, 2021

Kort om emnet

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 analyseres direkte. I disse tilfellene er god strukturering av rådata viktig.

*E-mail: martin.g.soyland@gmail.com

†E-mail: solveig.bjorkholt@stv.uio.no

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.

Hva lærer du?

Kunnskaper

Du vil lære:

- teknikker for effektiv samling av store mengder tekst fra offentlige kilder
- strukturere og klargjøre dokumentsamlinger for kvanitativ analyse
- bruksområder for tekst i statsvitenskaplig analyse
- analysemetoder for å trekke ut meningsfulle innsikter fra dokumentsamlinger

Ferdigheter

Du vil kunne:

- samle store mengder data fra strukturerte og ustrukturerte offentlige kilder
- klargjøre store dokumentsamlinger for analyse
- identifisere passende analyseteknikk for å besvare en problemstilling
- oppsummere og presentere analyseresultater av tekst på en forståelig måte

Generell kompetanse

Du vil kunne:

- gjøre utvalg og tilpasse kvantitative tekstanalyseteknikker for eget arbeid
- samle, strukturere, preprosessere og endre store mengder tekst
- anvende forskjellige analysemetoder og oppsummere resultater på en meningsfull måte
- utføre arbeidslivsrelevant tekstbehandling og analyse

Undervisning

Forelesning og seminar.

Forelesninger

Forelesningene vil dekke hovedtemaene i kurset med eksempler fra forskning og arbeidsliv. Tema for forelesning og seminar vil bli delt opp i hvordan å:

- anskaffe politiske tekster i forskjellige format
- strukturere tekster i form av kvantitative data
- oppsummere mengder tekst
- finne riktig analysemetode til tekstsamlinger for å svare på konkrete problemstillinger

Forslag til forelesninger:

1. Introduksjon

2. Anskaffelse og innlasting av tekst (a)
3. Bruke API – case: Stortinget
4. Forbehandling av tekst (a)
5. Forbehandling av tekst (b)
6. Veiledet læring (supervised learning) vs. ikke-veiledet læring (unsupervised learning)
7. Ordbøker og likhet mellom tekster
8. Klassifisering av tekst – temamodellering
9. Estimerer latent (ideologisk) posisjon fra tekst
10. Oppsummering

Seminarer

I seminarene vil studentene lære å utføre datasamling, strukturering av data og analyse av data, knyttet til forelesningene.

Seminarene vil fungere som et springbrett for deloppgavene over semesteret, der man går gjennom generelle trekk ved den aktuelle oppgaven.

Obligatoriske aktiviteter

- Oppmøte på seminarene
- Lovere oppgave etter hvert seminar

Eksamen

Semesteroppgave.

Semesteroppgaven bygges gradvis i løpet av seminarrekka. Delene samles til en fullstendig oppgave og leveres på slutten av semesteret.