Political Data Science

Lektion 6:

Tekst som data

Undervist af Jesper Svejgaard, foråret 2018
Institut for Statskundskab, Københavns Universitet
github.com/jespersvejgaard/PDS

I dag

1. Opsamling fra sidst
2. Dagens pensum
3. Workshop
4. Opsamling og næste gang

Overblik

1. Intro til kurset og R
2. R Workshop I: Explore
3. R Workshop II: Import, tidy, transform
4. R Workshop III: Programmering & Git
5. Web scraping & API
6. Tekst som data
7. Visualisering
8. GIS & spatiale data
9. Estimation & prædiktion
10. Superviseret læring I
11. Superviseret læring II
12. Usuperviseret læring
13. Refleksioner om data science
14. Opsamling og eksamen

Good thing we have DataCamp

Opsamling fra sidst

Opsamling fra sidst

Web & API

- API'er, der har klienter (pakker i R)
- API'er, hvor vi sender `GET()` og `POST()` forespørgsler med fx `httr`
- JSON-filer, arrays og objekter
- XML-filer og tags
- Web scraping med `rvest`

Dagens pensum

Hvorfor tekst som data?

(Grimmer & Stewart, 2013)

Problem:

.

Motivation:

.

.

Promise:

.

Overblik over metodiske tilgange

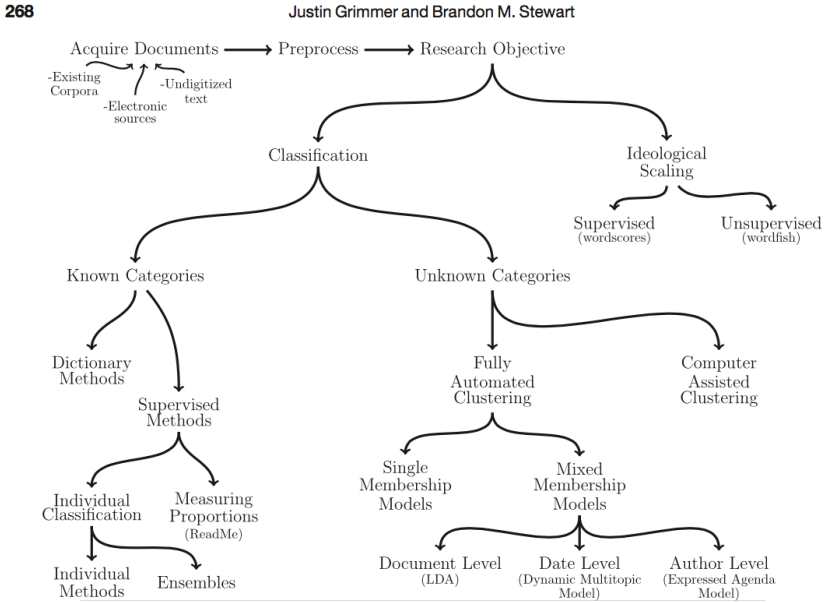


Fig. 1 An overview of text as data methods.

To hovedgrene inden for tekstanalyse

Klassifikation:

- Kendte kategorier - fx dictionary methods og superviseret klassifikation
- Ukendte kategorier - fx clustering

Skalering

- Kendt skala - fx word scores
- Ukendt skala - fx word fish

Klassifikation af tekster i kendte kategorier

Dictionary-metoder

Fremgangsmåde:

1. Kobl alle ordene i en tekst til et opslagsværk, fx over ordenes ladning (dikotom eller kontinuert)
2. Beregn tekstens gennemsnitlige score
3. Klassificér teksten
4. Validér: Sammenlign maskinkodning og menneskelig kodning

Antagelser:

- Ordenes score i opslagsværket matcher ordbrugen i dokumenterne.

Eksempel:

- Sammenligning af ladning i nyhedsmedier når der tales om klima (DataCamp)

Klassifikation af tekster i kendte kategorier

Superviseret læring

Fremgangsmåde:

1. Lav træningssæt: Mennesker kategoriserer et sample teksterne i et korpus (træningssæt)
2. Lær sammenhæng ml. features og kategorier: Sammenhæng ml. word count og kategorier
3. Validér: Cross-validation
4. Klassifikation: Resterende dokumenter klassificeres

Antagelser:

- Mønsteret i testsættet er lig med mønsteret i træningssættet (plausibelt v. random sampling)
- Ordene i et dokument er uafhængige af hinanden (ikke plausibelt, men virker)

Eksempel:

- Russisk militær-diskurs: Kodning af offentlige udtalelser fra civil og militær elite

Klassifikation af tekster i ukendte kategorier

Usuperviseret læring (FAC)

Formål:

- Inddel dokumenter i udtømmende og gensidigt udelukkende

Logik:

- Clusters har et centrum, fx gns. antal gange et bestemt ord optræder
- Hver tekst i en cluster har en afstand til et centrum, fx i form af summen af den kvadrerede forskellen ml. antal gange et ord optræder i en given tekst og antal gange ordet optræder i de andre tekster i clusteren
- Minimér summen af teksternes afstand til centrum i deres cluster ved i) at opdatere hvilken cluster en tekst er assigned til, ii) opdatere cluster centrum, iii) gentag

Skalering af tekster

Antagelser:

- Ideologisk dominans: Ideologisk udgangspunkt => ordbrug

Wordfish

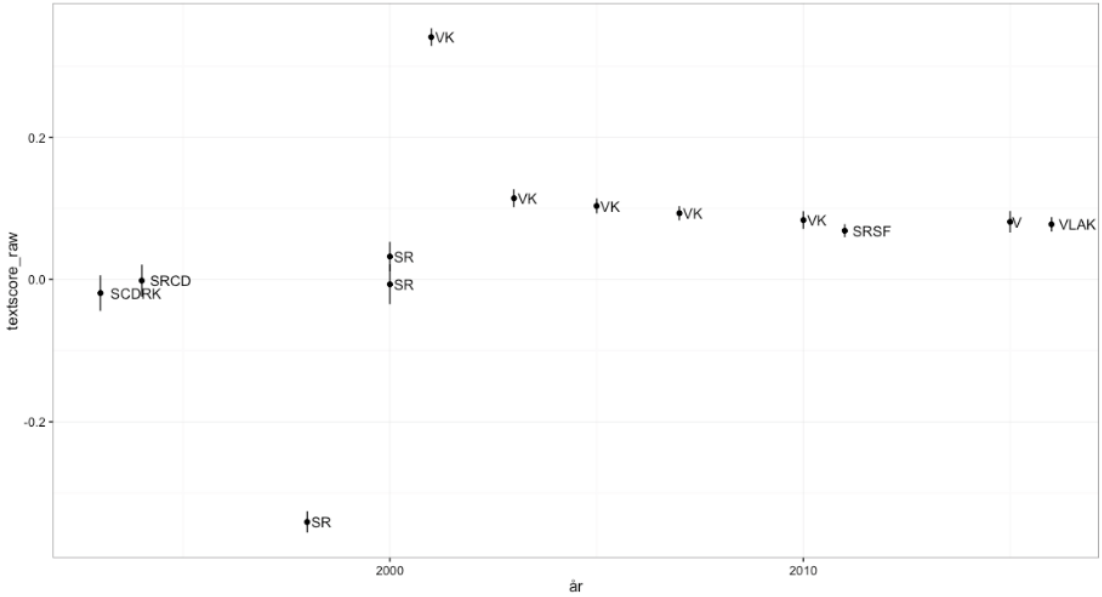
- Når skalaer er ukendte
- Svært at vide hvilken skala, man får ud - kan afspejle ikke-ideologisk skala

Wordscores:

- Når skalaer er kendte
- Bruger reference-tekster som skala
- Special case af dictionary metoder (beregner wordscores on the go, fx `tf_idf`)

Skalering af tekster

Eksempel



Typisk antagelse: Bag of words

Vi antager:

1. Ordenes rækkefølge er uden betydning.
2. Ordfrekvenser afspejler betydning.

Eksempel:

- "Vi ønsker hårdere straffe til kriminelle"
- "Vi ønsker ikke hårdere straffe til kriminelle men mildere straffe"

Funktionelle "løsninger":

- Antagelserne er kontroversielle og forkerte - men viser sig at være .
- Unigrams, bigrams og n-grams

Fire principper i automatiseret tekstanalyse

1. Alle modeller er forkerte, men nogen er brugbare
   - DGP for sprog er ukendt og kompliceret
2. Kvantitative metoder hjælper mennesker, men erstatter dem ikke
   - Mennesker guider analyse-processen, laver modellerne, fortolker og bruger output
   - Betydningen af kontekst og validering
3. Ingen metode er universelt bedst
   - Forskellige data => forskellige metoder
   - Forskellige forskningsspørgsmål => forskellige metoder
4. Validér, validér, validér
   - Når kategorier/skalaer er kendte => replikation og prædiktion
   - Når kategorier/skalaer er ukendte => validitet af koncepter

Typisk analyseprocess

1. Hent tekst-data
2. Præprocessering
   - fjern tal, tegn, kapitaler mm.
   - fjern stopord
   - fjern endelser (stemming)
   - fjern meget sjældne ord
3. Gør data tidy
   - konvertér fx til en (sparse) i x M document term matrix
4. Analyse

Typisk analyseproces ≠ universel analyseproces

Eksempel:

- Inferens af forfattere i Federalist Papers via "stopord" af Mosteller & Wallace (1963), gengivet i Imai (2017)

Pakker i R

`tm`

- text mining: import, præprocessering, konvertering til dtm

`quanteda`

- det samme, plus funktionalitet og ease

`stm`

- topic models

`stringr`

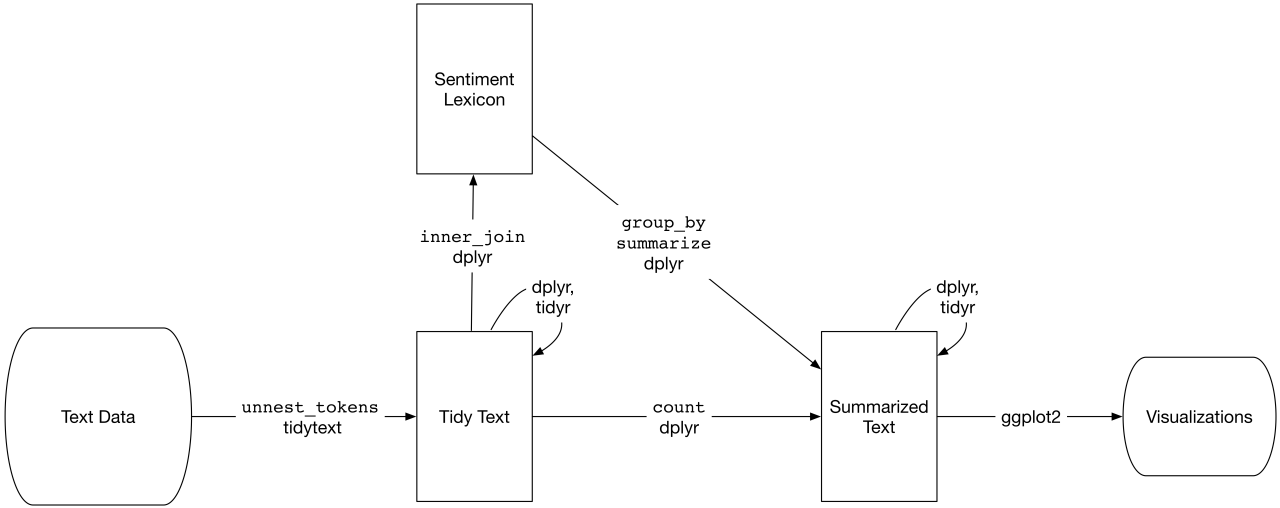
- tekst-manipulation

`tidytext`

- text mining: præprocessering og sentiment-analyse

Sentiment-analyse

Typisk workflow



Pakker med sentiments (danske og engelske)

```
# Loader sentiments fra "afinn" i `tidytext`
tidytext::get_sentiments("afinn")
```

```
# Loader danske sentiments fra nrc i `syuzhet`
syuzhet::get_sentiment_dictionary("nrc", language = "danish")
```

```
## # A tibble: 13,901 x 4
##      lang            word sentiment value
##     <chr>           <chr>     <chr> <dbl>
##  1 danish            abba  positive     1
##  2 danish            evne  positive     1
##  3 danish   nævnt ovenfor  positive     1
##  4 danish       absolutte  positive     1
##  5 danish syndsforladelse  positive     1
##  6 danish      absorberet  positive     1
##  7 danish        overflod  positive     1
##  8 danish         rigelig  positive     1
##  9 danish       akademisk  positive     1
## 10 danish         akademi  positive     1
## # ... with 13,891 more rows
```

Case: Køn journalistik

Case: Køn journalistik

Data:

- 1256 artikler fra The Guardian om sexual harassment fra før/efter Weinstein

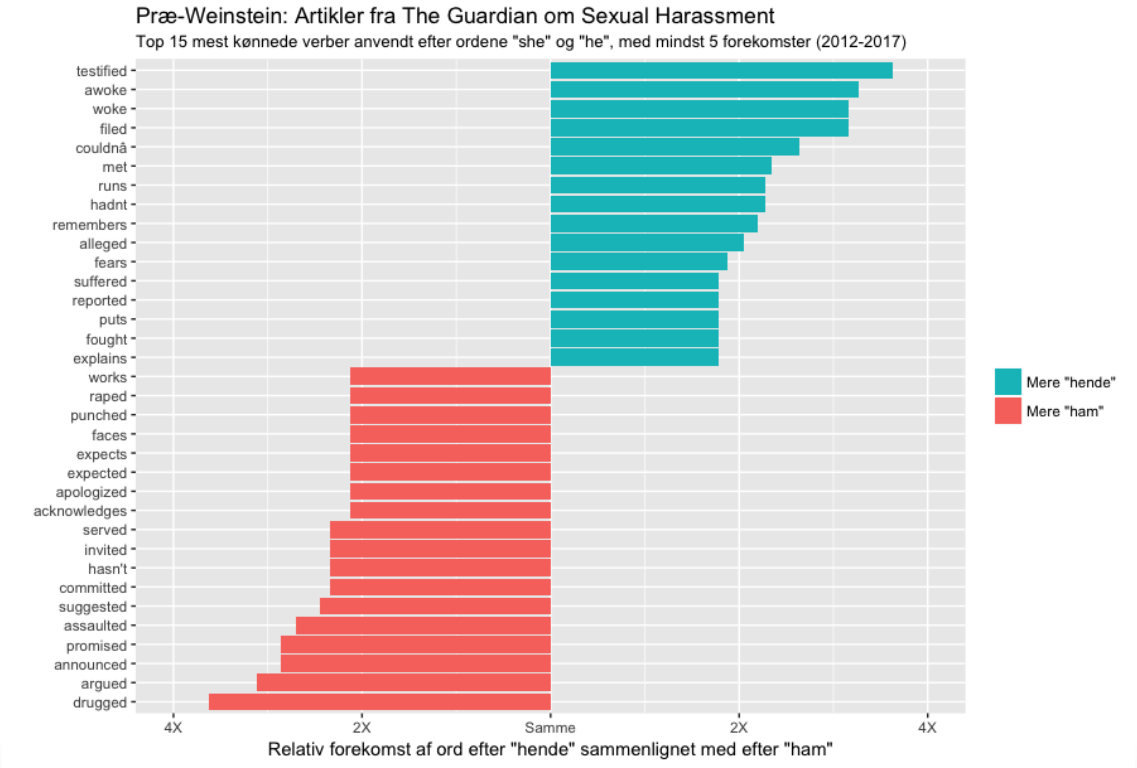
Tilgang:

- Hente tekst (API)
- Præprocessere tekst-data (`dplyr`, `tidyr`, `stringr`, `rvest`, loops)
- Analysere (`dplyr`, `magrittr`)
- Visualisere (`ggplot2`)

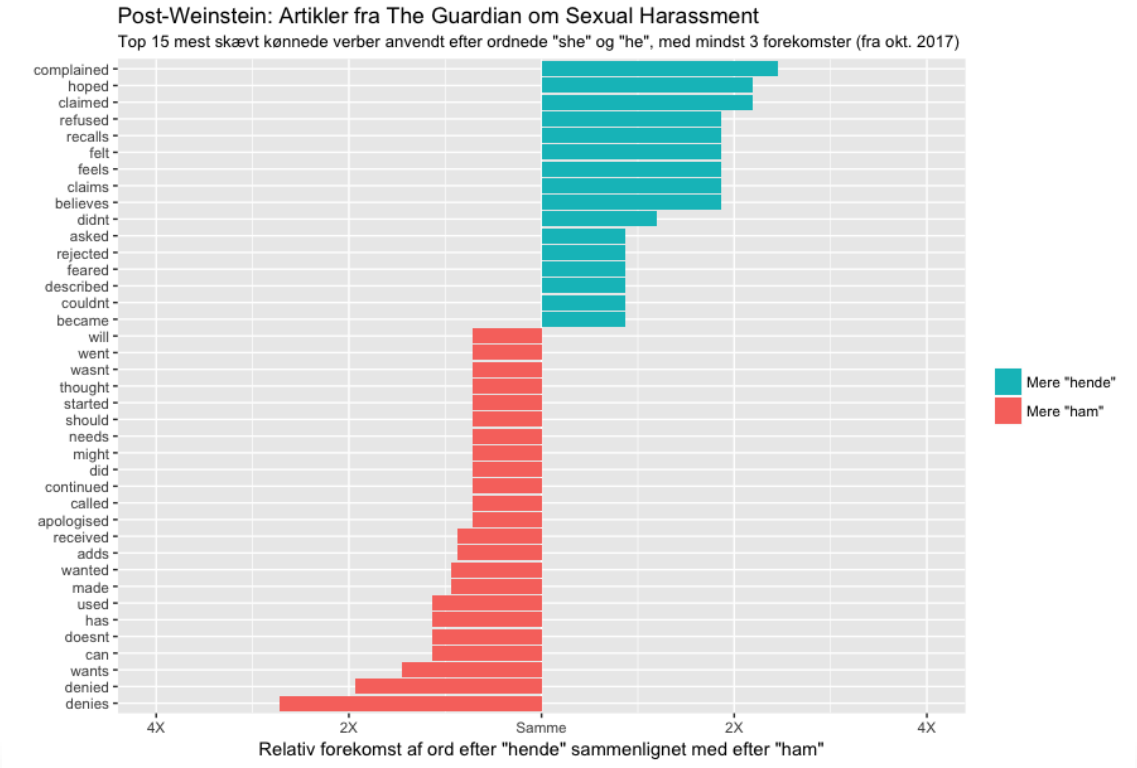
Forskningsspørgsmål?

- Er der forskel på hvilke ord, som typisk kommer efter hhv. "he" og "she"?
- Er der forskel på ovenstående før/efter Weinstein skandalen breakede?

Case: Køn journalistik



Case: Køn journalistik



Opgave-session

Opgave-session

Opgave 1

Hent datasættet `bigrams.csv` fra PDS/data/ på GitHub. Bigrammerne er baseret på alle artikler fra The Guardian i perioden 01-01-2013 til 01-01-2018 der matcher søgeordene "sexual harassment". Kolonnen "articles" angiver, om bigrammet stammer fra en artikel udgivet før/efter Weinstein-skandalen breakede 5. oktober 2017. Lav en sentiment-analyse af de ord, som henholdsvist følger efter "he" og "she". Er der forskel før/efter Weinstein-skandalen? Er der forskel generelt? Visualiser dine indsigter.

Brug fx pakken `tidytext` (hint: `get_sentiments()` og `inner_join()`).

Opgave 2

Hent dit eget korpus af tekster fra The Guardian via deres API. Du kan registrere dig til en API key via linket her. Brug API-klienten i pakken `GuardianR`. Find evt. inspiration og hjælp til at hente og præprocessere data i scriptet `06_tekst_pre.R` på fagets GitHub. Sæt en periode og hent artikler pba. et eller flere selvvalgte søgeord. Definer en sentiment-problemstilling og belys den.

Opsamling og næste gang

Vigtigste pointer fra i dag

- Klassifikation og skalering af tekster
- Superviserede og usuperviserede metoder
- Præprosessering af tekst

Næste gang

- Indhold:
  - Visualisering
- Pensum:
  - DVSS: "Before you begin"
  - DVSS: kap 3 + 8
  - CS: Visualisation
- DataCamp:
  - Data Visualization in R
- Supplerende DataCamp, hvis man keder sig:
  - Data Visualization with ggplot2 (Part 1)
  - Data Visualization with ggplot2 (Part 2)
  - Data Visualization with ggplot2 (Part 3)

Tak for i dag!