Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

# **Tekstmining:**
# En kort innføring

Pierre Lison
*Seniorforsker*, NR

Norsk evalueringsforening

30.08.2018

# Litt om NR

- Et forskningsinstitutt (privat stiftelse)
- Utfører **oppdragsforskning** for både næringsliv og offentlig sektor, i Norge og internasjonalt
- Fagområder: Anvendelse av statistisk modellering, maskinlæring, mønstergjenkjenning og IKT
- Vertsinstitusjon for *BigInsight* (senter for forskningsdrevet innovasjon) som tar sikte på å utvikle analytiske verktøy for å trekke ut kunnskap fra komplekse datasett

# Tekstmining


- «*Vi drukner i data, men tørster etter informasjon*» (Thomas H. Eriksen)
- Mesteparten av dataene vi behandler er i **tekstform** (dokumenter, eposter, nettsider, osv.)

# Tekstmining


- Tekstmining handler om å finne **verdifull informasjon** i store *mengder tekst*
- ≠ fra «klassiske» *data mining*, som forsøker å avdekke mønstre i *strukturerte data* (tabeller osv.)

# Utfordringer

- Tekster er mye mer komplisert å «forstå» (for en maskin) enn tabeller
- Språk er iboende **flertydig**
  - f.eks. «bank» (en finansinstitusjon eller et fiskerikt område?)
  - Kontekst er viktig!
- Mange måter å utrykke den samme idéen
  - Synonymer, alternative formuleringer, språkvarianter, osv.

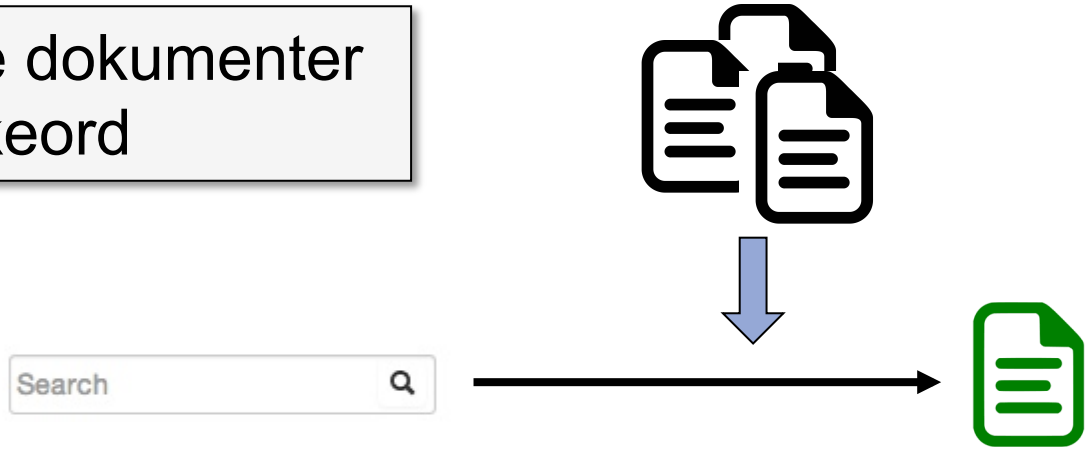
# Oppgaver i tekstmining

- Søk
- Tekstklassifisering
- Informasjonsekstraksjon
- Sentimentanalyse

# Oppgaver i tekstmining

- **Søk**
- Informasjonsekstraksjon
- Tekstklassifisering
- Sentimentanalyse

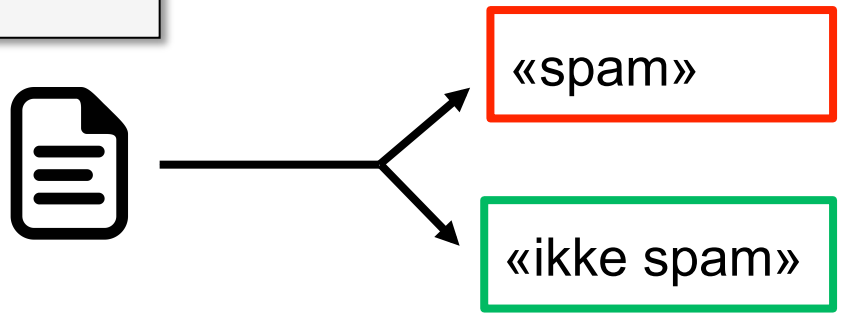
**Mål**: finne relevante dokumenter gitt et eller flere søkeord



# Oppgaver i tekstmining

- Søk
- Informasjonsekstraksjon
- **Tekstklassifisering**
- Sentimentanalyse

**Mål**: automatisk fordele dokumenter i bestemte kategorier



# Oppgaver i tekstmining

- Søk
- Informasjonsekstraksjon
- **Tekstklassifisering**
- Sentimentanalyse

**Clustering:** automatisk gruppere dokumenter basert på likhetstrekk (f.eks. *topic modelling*)

# Oppgaver i tekstmining

- Søk
- Tekstklassifisering
- **Informasjonsekstraksjon**
- Sentimentanalyse

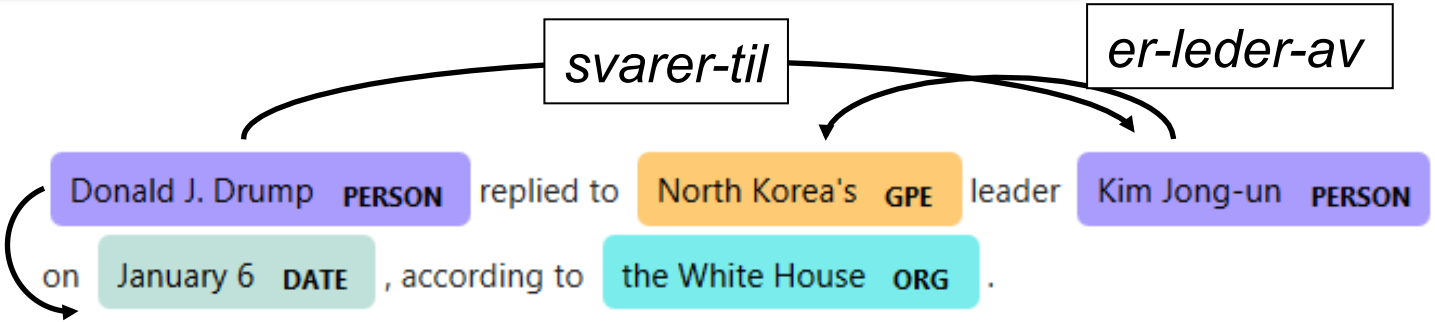
**Gjenkjenning av navngitte enheter:** Finne forekomster av personer, bedrifter, steder, osv. i tekst

Donald J. Drump PERSON replied to North Korea's GPE leader Kim Jong-un PERSON on January 6 DATE , according to the White House ORG .

# Oppgaver i tekstmining

- Søk
- Tekstklassifisering
- **Informasjonsekstraksjon**
- Sentimentanalyse

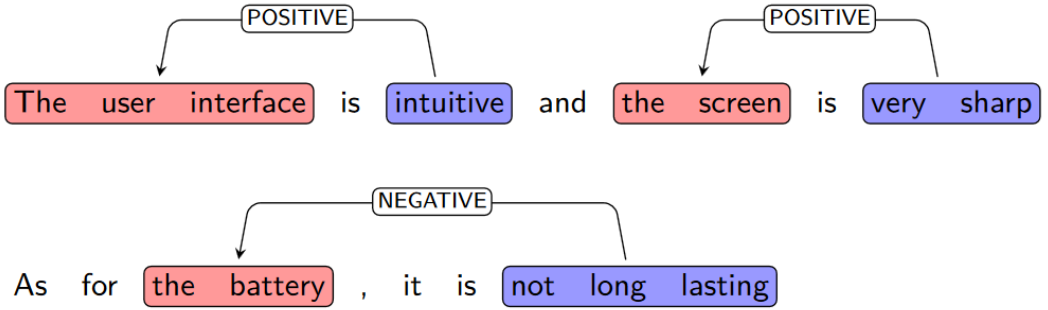
Vi kan også utvikle verktøy til å automatisk ekstrahere **relasjoner** mellom enhetene



# Oppgaver i tekstmining

- Søk
- Informasjonsekstraksjon
- Tekstklassifisering
- **Sentimentanalyse**

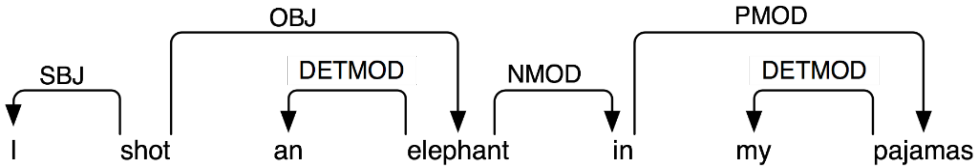
**Mål**: Identifisere *subjektive meninger* (følelser osv.) som uttrykkes i tekst.



# Tekstminingskomponenter

- Tekstsegmentering (avsnitt, setninger, ord)
- Ordklassetagging
- Parsing (gjenkjenning av setningsstruktur)

… og mange flere




Typisk basert på **maskinlæringsmodeller** trent på store mengde (annoterte) tekstdata

# Evaluering

- Vi evaluerer ytelsen til tekstminingmodeller på *testdata* (som er isolert fra treningsdata)
- To typer feil:
  - Falske positiver
  - Falske negativer

Output fra systemet

Sanne verdi

| | x | v |
|---|---|---|
| x | Sanne negative | Falske positive |
| v | Falske negative | Sanne positive |

- *Presisjonen* og *treffraten* kan beregnes ut fra disse

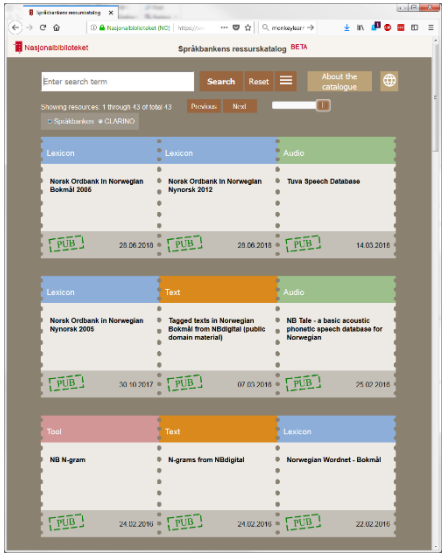
# I praksis

- Hvordan kan jeg bruke tekstmining på mine dokumenter?
- For «store» språk og tekstsjangre er det allerede IT-verktøy som kan gjøre jobben fos oss
- Men ofte (spesielt for norske tekster) må man utvikle verktøyene selv, ved å:
  - Samle og annotere tekstdata
  - Trene maskinlæringsmodeller på disse

# Språkressurser i norsk

- **Språkbanken** fra Nasjonalbiblioteket tilbyr en rekke språkressurser
  - Tekstsamlinger, ordbank, taledatabaser, oversettelser, osv.
  - Både bokmål og nynorsk
- Men fremdeles langt bak «store» språk som engelsk

Nasjonalbiblioteket



# Forskningsfronten

- Mining i dokumenter som blander flere medier (tekst, bilde, lyd, video, osv.)


- Flerspråklige modeller
- Informasjonsekstraksjon med liten/ingen annoterte data
- Neuralske nettverker («deep learning»)

# Spørsmål?


- Kontakt: plison@nr.no