Computationel Tekstanalyse i Samfundsvidenskaberne

Social Data Science I

18. oktober 2023

Kristian Gade Kjelmann
kgk@socsci.aau.dk

Inst. for Sociologi og Socialt Arbejde
Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Dagens program

- Introduktion
  - Læringsmål for i dag
  - Hvorfor er computationel tekstanalyse interessant?
  - NLP eller computationel tekstanalyse?
- Brug af computationel tekstanalyse i samfundsvidenskaberne
  - Computationel tekstanalyse illustreret
  - Eksempler på brug af computationel tekstanalyse
- Introduktion til Natural Language Processing
- Introduktion til sprogmodeller
- Brug af spaCy i Python (live-coding)
- Brug af dictionary metoder og sentiment analyse i Python (live-coding)
- Teori i computationel tekstanalyse
  - Udvælgelse af corpus
  - Valg af metode og modeller
  - Fortolkning og validering

Læringsmål for i dag
Introduktion

- Have kendskab til centrale koncepter inden for "computationel tekstanalyse".

Læringsmål for i dag
Introduktion

- Have kendskab til centrale koncepter inden for "computationel tekstanalyse".
- Forståelse for generelle muligheder med computationel tekstanalyse og natural language processing.

Læringsmål for i dag
Introduktion

- Have kendskab til centrale koncepter inden for "computationel tekstanalyse".
- Forståelse for generelle muligheder med computationel tekstanalyse og natural language processing.
- Forståelse for væsentlige udfordringer i at anvende computationel tekstanalyse til samfundsvidenskabelige analyser.

- Have kendskab til centrale koncepter inden for "computationel tekstanalyse".
- Forståelse for generelle muligheder med computationel tekstanalyse og natural language processing.
- Forståelse for væsentlige udfordringer i at anvende computationel tekstanalyse til samfundsvidenskabelige analyser.
- Forståelse for teknikker anvendt i natural language processing.

Læringsmål for i dag
Introduktion

- Have kendskab til centrale koncepter inden for "computationel tekstanalyse".
- Forståelse for generelle muligheder med computationel tekstanalyse og natural language processing.
- Forståelse for væsentlige udfordringer i at anvende computationel tekstanalyse til samfundsvidenskabelige analyser.
- Forståelse for teknikker anvendt i natural language processing.
- Kompetence til at anvende eksisterende sprogmodeller i Python.

Læringsmål for i dag
Introduktion

- Have kendskab til centrale koncepter inden for "computationel tekstanalyse".
- Forståelse for generelle muligheder med computationel tekstanalyse og natural language processing.
- Forståelse for væsentlige udfordringer i at anvende computationel tekstanalyse til samfundsvidenskabelige analyser.
- Forståelse for teknikker anvendt i natural language processing.
- Kompetence til at anvende eksisterende sprogmodeller i Python.
- Kompetence til at anvende dictionary model på tekstdata i Python.

Muligheder med data

- Øget tilgængelighed af eksisterende tekstdata (arkiver, nyhedsmedier, policy dokumenter)
- Tidligere produceret kvalitative data (interviewtransskriberinger, lydoptagelser)
- Øget mængde af "digitalt fødte"data (sociale medier, kommentarspor, anmeldelser, reaktioner)

Muligheder med metoder

- Bedre til at udnytte at data er digitale
- Tilgængelighed af hardware tillader bearbejdning af større mængder af data
- Øget genbrug af tilgængelige ressourcer (sprogmodeller, dictionaries, trænede modeller og classifiers)

Natural Language Processing (NLP)

Computervidenskabelig disciplin, der beskæftiger sig med, hvordan en computer kan forstå og producere menneskeligt sprog

Computationel tekstanalyse ("text mining")

Anvendelse af natural language processing til analytiske formål.

- NLP teknikker ofte mere avancerede og detaljerede, end hvad der er nødvendigt for samfundsvidenskabelig analyse (Stuhler, 2022)
- Semantisk frem for syntaktisk fokus (Franzosi, 1989)

Samfundsforskeren

- Ofte interesseret i latente koncepter
- Må acceptere mangel på valideringsdata
- Interesseret i specifik kontekst
- Skepsis ift. menneskelige udsagn og vurdering

(Maskinlærings)ingeniøren

- Ofte interesseret i målbare koncepter
- Forventer valideringsdata
- Interesseret i generisk kontekst (genanvendelighed)
- Accepterer menneskelige udsagn og vurdering (data er data)

- Skabe overblik (fx nøgleordsanalyse)

Computationel tekstanalyse i samfundsvidenskaberne

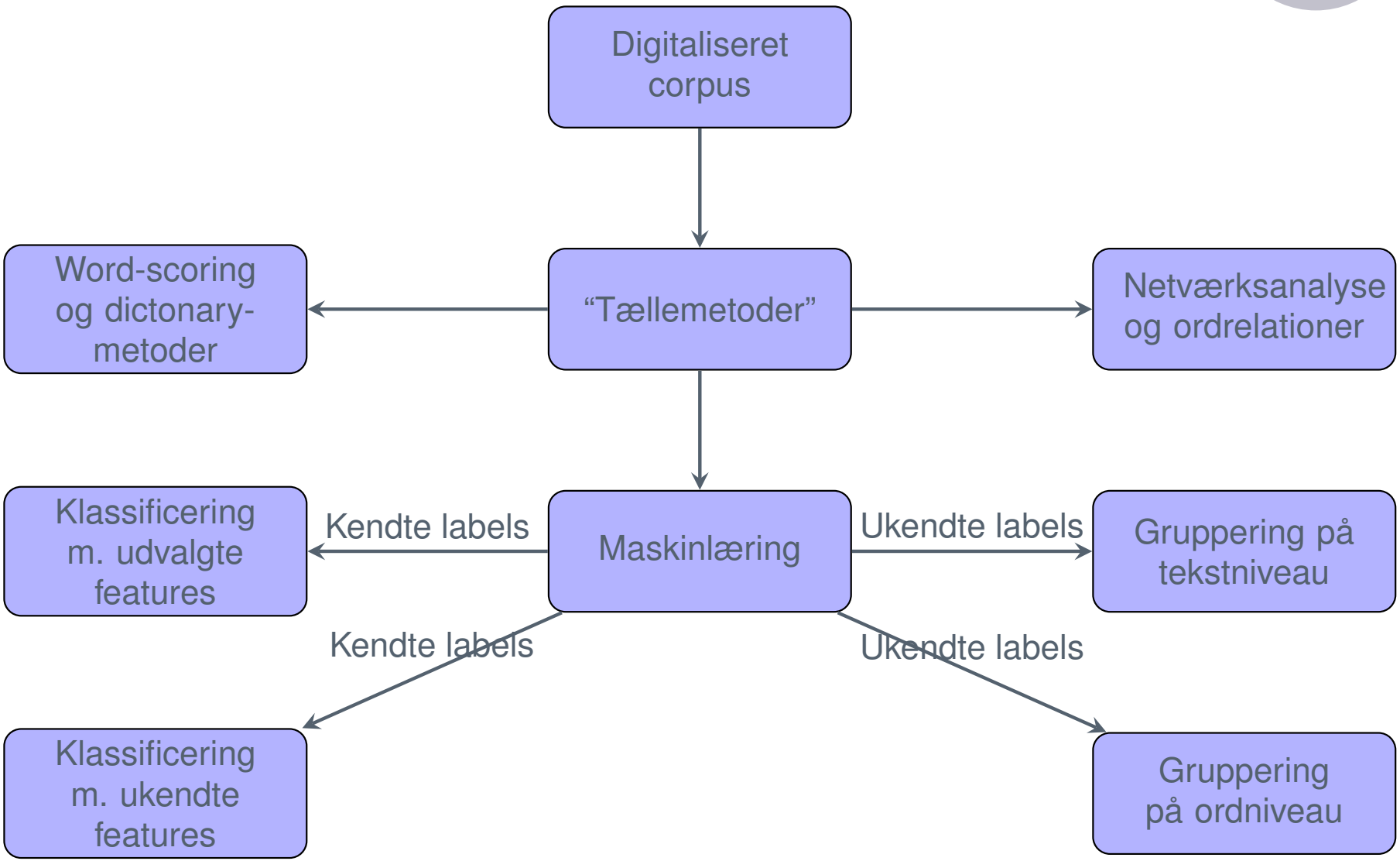
- Skabe overblik (fx nøgleordsanalyse)
- Identificér og måle prædefinerede, relevante koncepter (hate-speech, politisk ideologi, diskurs)
  - Udlede metrikker/variable

Computationel tekstanalyse i samfundsvidenskaberne

- Skabe overblik (fx nøgleordsanalyse)
- Identificér og måle prædefinerede, relevante koncepter (hate-speech, politisk ideologi, diskurs)
  - Udlede metrikker/variable
- Udforske og forstå komplekse meningssammenhænge (hvordan temaer opstår og udvikler sig, hvordan befolkningsgrupper, institutioner eller andet italesættes, koblinger mellem temaer, holdning og mening)

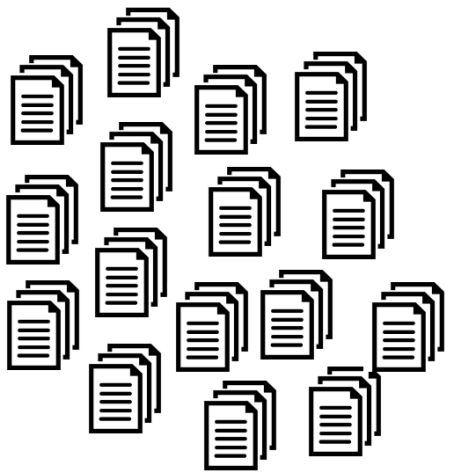
Computationel tekstanalyse i samfundsvidenskaberne

- Skabe overblik (fx nøgleordsanalyse)
- Identificér og måle prædefinerede, relevante koncepter (hate-speech, politisk ideologi, diskurs)
  - Udlede metrikker/variable
- Udforske og forstå komplekse meningssammenhænge (hvordan temaer opstår og udvikler sig, hvordan befolkningsgrupper, institutioner eller andet italesættes, koblinger mellem temaer, holdning og mening)
- Identificér sociale aktører og deres (formodede) handlinger (hvem gjorde hvad til hvem?)
  - Brug af "dependency parsing"(computationel analyse af sætningskonstruktion)


Digitaliseret corpus
Word-scoring og dictonary-metoder
"Tællemetoder"
Netværksanalyse og ordrelationer
Klassificering m. udvalgte features
Kendte labels
Maskinlæring
Ukendte labels
Gruppering på tekstniveau
Kendte labels
Ukendte labels
Klassificering m. ukendte features
Gruppering på ordniveau

Fra tekst til corpus

Computationel tekstanalyse illustreret



Så vil ministeren garantere, at der inden for få dage - bliver det jo nødt til at være, fordi vi jo står over for et månedsskifte - eller i hvert fald kort tid vil kunne findes en løsning, der sikrer, at amu-centrene ikke behøver at fyre medarbejdere?
Ministeren.
Det der med at love, hvor hurtigt man kan gøre tingene i en tid, hvor alt gøres på fuld speed, vil jeg simpelt hen ikke. Jeg vil gerne love, at jeg gør tingene så hurtigt, jeg overhovedet kan slippe af sted med det, og jeg mener sådan set, at det er rigtigt at gå ind og begynde at kigge på, hvad vi kan stille op i forhold til amu-sektoren.
Jeg vil gerne sige, at jeg mener, at der er nogle udfordringer forbundet med det. En af udfordringerne og også årsagen til, at jeg var lidt tøvende indledningsvis og i starten, er jo, at jeg som minister har et håb om, at vi, hvad angår den krise, vi desværre nok ser ind i, kommer klogere ud af den, end vi gjorde, sidst der var en krise, og det vil altså sige med langt større tryk på og mulighed for at tage efteruddannelse. Og dermed har jeg jo et håb om, at amu-sektoren kommer til at blomstre. For det er nogle af dem, der skal levere noget af den uddannelse, der skal til. Det vil jo sige, at vi her taler om noget, hvor det helst skal være sådan, at man skal tørskoet gennem i en kort periode, men at vi så faktisk helst skal se, at amu-sektoren virkelig blomster og jo dermed også får indtægter ind.
Det er desværre på ryggen af en krise, men det er faktisk nogle af dem, der skal hjælpe os igennem på en måde, så vi også kommer klogere ud på den anden side. Så det har været årsagen til den indledningsvise tøven.
Spørgeren.
Det er jo netop en begrundelse for at hjælpe den her sektor, så vi har den sektor, når vi får brug for den. Men i forlængelse af det vil jeg så høre ministeren, om hun også har fået kigget på andre former for skoler, der har indtægtsdækket virksomhed, og om hun også er kommet frem til, at de har brug for hjælp, eller om det kun drejer sig om amu-sektoren.
Ministeren.
Det synes jeg faktisk er mere dilemmafyldt end som så, for - for at sige det en lille smule banalt - i hvor høj grad skal man lave situationen sammenlignelig med resten af det private erhvervsliv for de ting, der faktisk er privat? Der mener jeg egentlig, at der skal være en ret høj grad af parallelitet. Derfor skal man finde nogle veje. Jeg synes ikke, vi har fundet den rigtige vej på den del, der handler om indtægtsdækket virksomhed. Derfor har jeg egentlig også lyst til at sige, at jeg gerne vil drøfte amu, men når det gælder det andet, har jeg simpelt hen ikke - for nu at sige det ligeud - fundet de vises sten, så vi kan undgå, at der ikke er urimelige forhold i forhold til andre dele af erhvervslivet. Så det er i min verden to forskellige ting. Jeg har været mægtig glad for, at spørgeren så ihærdigt har taget det op. Det tror jeg driver os frem mod måske at finde en løsning på et tidspunkt. Det er bare for at være ærlig om, hvad jeg synes dilemmaet er.
Spørgeren.
Jeg tror egentlig, at vi er kommet rundt om det, men jeg glæder mig til, at vi, sådan som jeg kan forstå det, i løbet af meget kort tid får løsningen. Det håber jeg er i løbet af en uges tid eller noget i den stil, for det her haster virkelig. Tak for svarene.
Ministeren.

Fra corpus til concordance

Computationel tekstanalyse illustreret

t jeg kan se , at det er kommuner med en høj andel af indvandrere og efterkommere af indvandrere , hvor smittetallene s
er med en høj andel af indvandrere og efterkommere af indvandrere , hvor smittetallene stiger og stiger . Jeg vil bare
ren , at regeringens målsætning om at få 10.000 flere indvandrere i arbejde er lykkedes ? Værsgo . Tak for det . Jeg vi
ner , at regeringens målsætning om at få 10.000 flere indvandrere i arbejde er lykkedes ? Ministeren . Det er rigtigt ,
ingen har et ambitiøst mål om , at beskæftigelsen for indvandrere og efterkommere skal øges med 10.000 personer frem mo
, at der er alt for mange med ikkevestlig baggrund , indvandrere og efterkommere , som rent faktisk står uden for arbe
det ganske kort : I 2008 havde vi 68.000 ikkevestlige indvandrere uden for arbejdsmarkedet , og det tal var i 2013 steg
V):Mener ministeren , at beskæftigelsesfrekvensen for indvandrere og flygtninge i den arbejdsdygtige alder er høj nok ?
: Mener ministeren , at beskæftigelsesfrekvensen for indvandrere og flygtninge i den arbejdsdygtige alder er høj nok ?
le . Tidligere har det været sådan , at flygtninge og indvandrere faktisk selv gik og sagde , at det var økonomisk attr
fortælle , at der i 2001 i Danmark var 14.111 ledige indvandrere / efterkommere fra ikkevestlige lande . Tallet stiger
nisteren har haft ansvaret for beskæftigelsen også af indvandrere / efterkommere . Der er jo ingen tvivl om , at alle s
, at man vælger at give ekstra penge til arbejdsløse indvandrere , som kan opnå 2.000 kr. ekstra skattefrit om måneden
tholde . Man har forhøjet ydelserne til udlændinge og indvandrere , man har lempet kravene for optjening af dagpenge ti
ennesker på pension og hæve ydelserne til arbejdsløse indvandrere . Derfor er mit spørgsmål til Socialdemokraternes ord
opbakning til ligestilling mellem kønnene , både hos indvandrere og hos efterkommere og hos etniske danskere . Man kun
om den måde , man ser på flygtninge på , og vi ser på indvandrere på , men her spørger jeg principielt set : Er det en
islamiske værdier , og hvor mange af de uintegrerbare indvandrere bor . Jeg formoder , at det gav bonus på valgnatten ,
fortæller , at muslimer er sådan og sådan , eller at indvandrere og efterkommere ikke har fortjent det samme som alle
rældersyet , som jeg siger , til at ramme de såkaldte indvandrere . Det er jo sådan lidt blevet et begreb for alt mulig
-timersreglen og andre fattigdomsydelser , som holder indvandrere og alle andre fast i store problemer , og som forværr
r nylig sagde , at ud over nogle få grønthandlere har indvandrere ikke bidraget med noget . Altså , det betyder jo noge
ørselsindkomst , fordi det - og jeg citerer - går til indvandrere og efterkommere . Jeg skal bare lige have opklaret ,
f.eks. folkepensionister ? Der vil jo være masser af indvandrere , som bor i Danmark , som har knoklet i Danmark hele
sindkomst , fordi det også går til - og jeg citerer - indvandrere og efterkommere . Nu siger ordføreren , at det skal k
dtalelse om , at det også skulle være efterkommere og indvandrere . Det er trods alt godt at høre . Det er mærkeligt ,
r kvinder ? Ordføreren . I forhold til spørgsmålet om indvandrere og efterkommere er det jo typisk sådan , at udlænding
, men ellers i statistikkerne bliver registreret som indvandrere og efterkommere . Så er der nogle af dem , der har få
p 20 år . Jeg har arbejdet på skoler med rigtig mange indvandrere , og det er på deres vegne , jeg føler mig dybt kræn
Ordføreren siger : Jeg har været skolelærer for mange indvandrere . Indvandrere , altså drenge og piger , som er født o
er : Jeg har været skolelærer for mange indvandrere . Indvandrere , altså drenge og piger , som er født og opvokset i D

Tællemetoder: Hyppigste ord

Computationel tekstanalyse illustreret

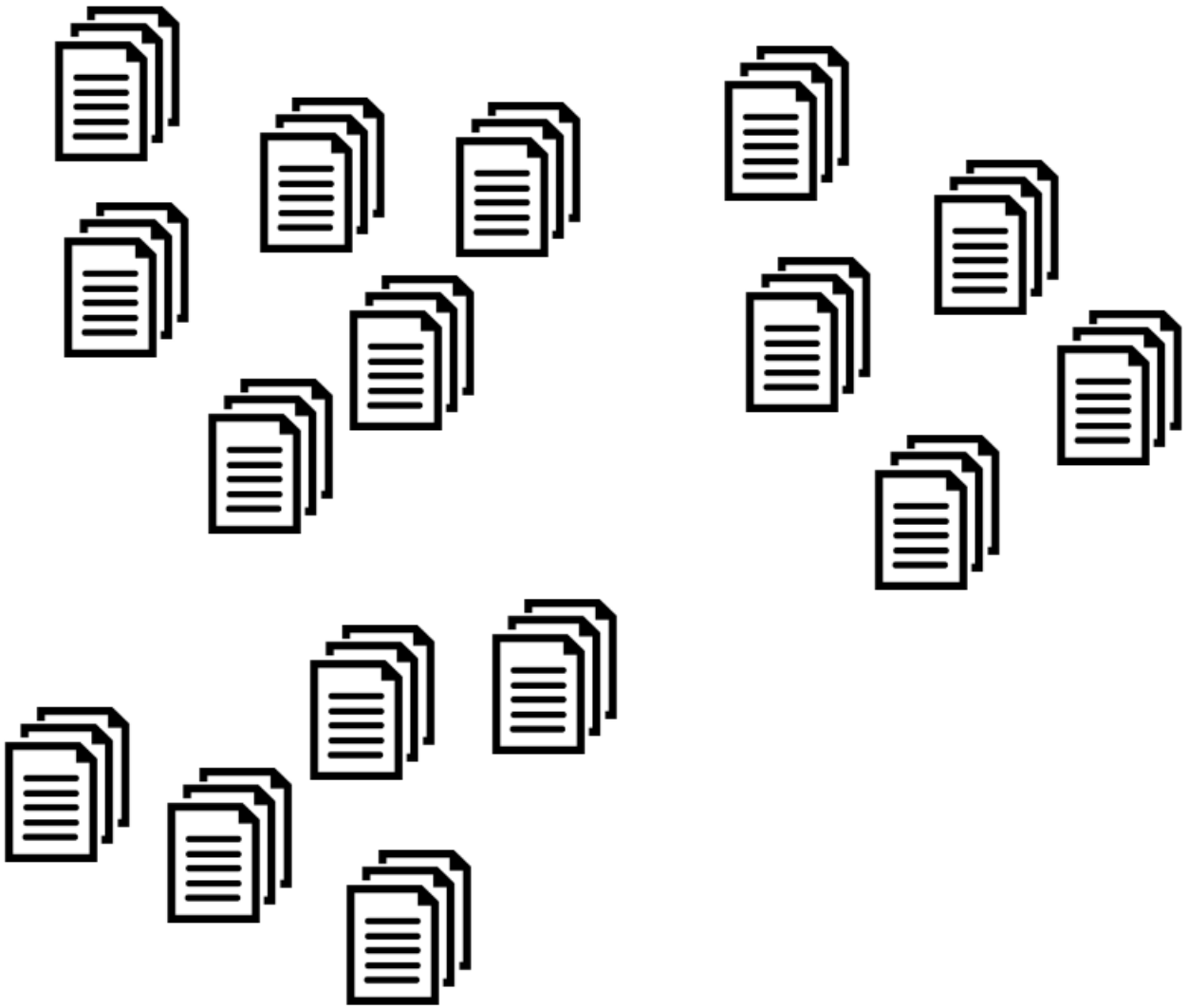
t jeg kan se , at det er kommuner med en høj andel af indvandrere og efterkommere af indvandrere , hvor smittetallene s
er med en høj andel af indvandrere og efterkommere af indvandrere , hvor smittetallene stiger og stiger . Jeg vil bare
ren , at regeringens målsætning om at få 10.000 flere indvandrere i arbejde er lykkedes ? Værsgo . Tak for det . Jeg vi
ner , at regeringens målsætning om at få 10.000 flere indvandrere i arbejde er lykkedes ? Ministeren . Det er rigtigt ,
ingen har et ambitiøst mål om , at beskæftigelsen for indvandrere og efterkommere skal øges med 10.000 personer frem mo
, at der er alt for mange med ikkevestlig baggrund , indvandrere og efterkommere , som rent faktisk står uden for arbe
det ganske kort : I 2008 havde vi 68.000 ikkevestlige indvandrere uden for arbejdsmarkedet , og det tal var i 2013 steg
V):Mener ministeren , at beskæftigelsesfrekvensen for indvandrere og flygtninge i den arbejdsdygtige alder er høj nok ?
: Mener ministeren , at beskæftigelsesfrekvensen for indvandrere og flygtninge i den arbejdsdygtige alder er høj nok ?
le . Tidligere har det været sådan , at flygtninge og indvandrere faktisk selv gik og sagde , at det var økonomisk attr
fortælle , at der i 2001 i Danmark var 14.111 ledige indvandrere / efterkommere fra ikkevestlige lande . Tallet stiger
nisteren har haft ansvaret for beskæftigelsen også af indvandrere / efterkommere . Der er jo ingen tvivl om , at alle s
, at man vælger at give ekstra penge til arbejdsløse indvandrere , som kan opnå 2.000 kr. ekstra skattefrit om måneden
tholde . Man har forhøjet ydelserne til udlændinge og indvandrere , man har lempet kravene for optjening af dagpenge ti
ennesker på pension og hæve ydelserne til arbejdsløse indvandrere . Derfor er mit spørgsmål til Socialdemokraternes ord
opbakning til ligestilling mellem kønnene , både hos indvandrere og hos efterkommere og hos etniske danskere . Man kun
om den måde , man ser på flygtninge på , og vi ser på indvandrere på , men her spørger jeg principielt set : Er det en
islamiske værdier , og hvor mange af de uintegrerbare indvandrere bor . Jeg formoder , at det gav bonus på valgnatten ,
fortæller , at muslimer er sådan og sådan , eller at indvandrere og efterkommere ikke har fortjent det samme som alle
rræddersyet , som jeg siger , til at ramme de såkaldte indvandrere . Det er jo sådan lidt blevet et begreb for alt mulig
-timersreglen og andre fattigdomsydelser , som holder indvandrere og alle andre fast i store problemer , og som forværr
r nylig sagde , at ud over nogle få grønthandlere har indvandrere ikke bidraget med noget . Altså , det betyder jo noge
ørselsindkomst , fordi det - og jeg citerer - går til indvandrere og efterkommere . Jeg skal bare lige have opklaret ,
f.eks. folkepensionister ? Der vil jo være masser af indvandrere , som bor i Danmark , som har knoklet i Danmark hele
sindkomst , fordi det også går til - og jeg citerer - indvandrere og efterkommere . Nu siger ordføreren , at det skal k
dtalelse om , at det også skulle være efterkommere og indvandrere . Det er trods alt godt at høre . Det er mærkeligt ,
r kvinder ? Ordføreren . I forhold til spørgsmålet om indvandrere og efterkommere er det jo typisk sådan , at udlænding
, men ellers i statistikkerne bliver registreret som indvandrere og efterkommere . Så er der nogle af dem , der har få
p 20 år . Jeg har arbejdet på skoler med rigtig mange indvandrere , og det er på deres vegne , jeg føler mig dybt krænk
Ordføreren siger : Jeg har været skolelærer for mange indvandrere . Indvandrere , altså drenge og piger , som er født o
er : Jeg har været skolelærer for mange indvandrere . Indvandrere , altså drenge og piger , som er født og opvokset i D

Dictionary-metoder: Udled specifikke ord

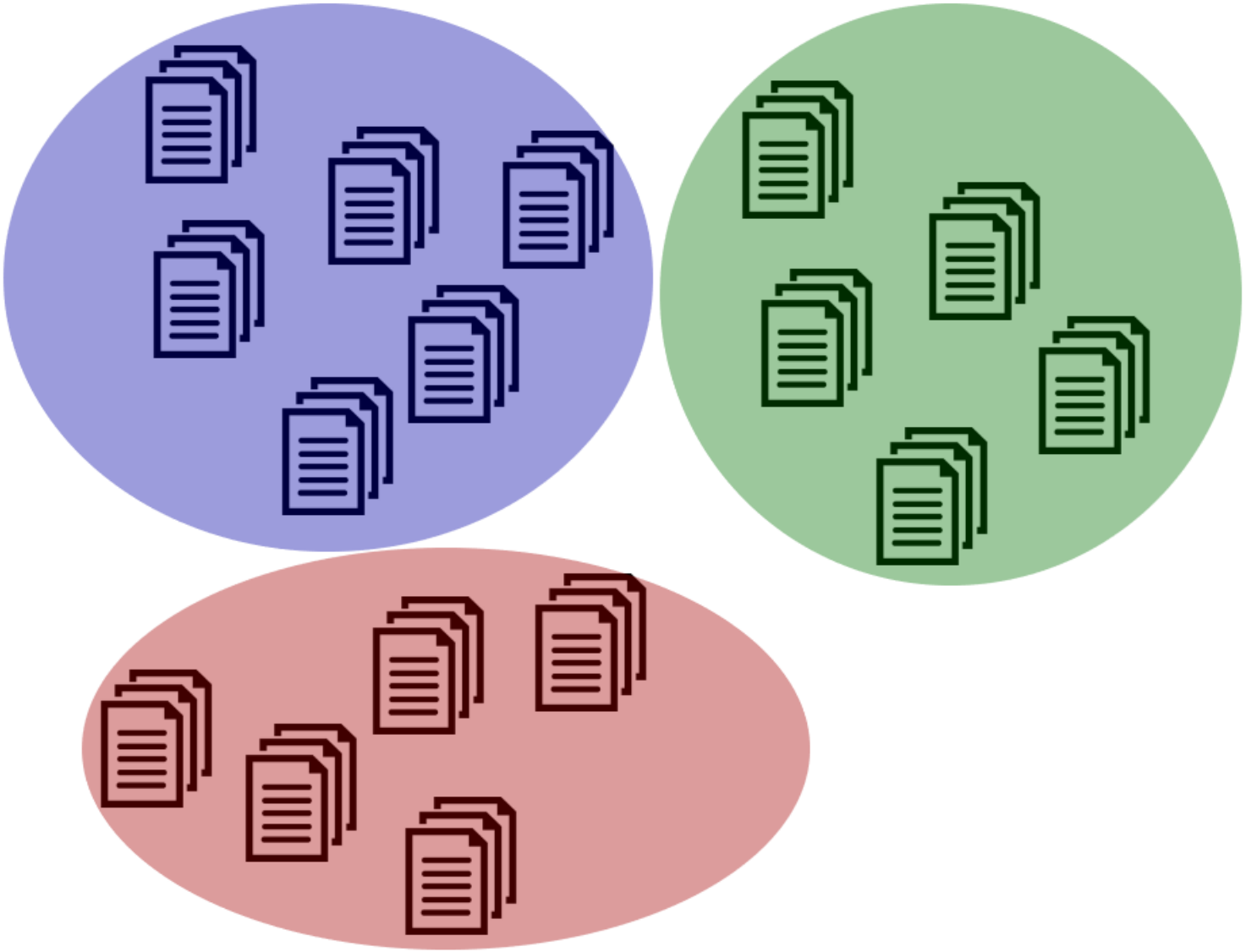
Computationel tekstanalyse illustreret

t jeg kan se , at det er kommuner med en høj andel af indvandrere og efterkommere af indvandrere , hvor smittetallene s
er med en høj andel af indvandrere og efterkommere af indvandrere , hvor smittetallene stiger og stiger . Jeg vil bare
ren , at regeringens målsætning om at få 10.000 flere indvandrere i arbejde er lykkedes ? Værsgo . Tak for det . Jeg vi
ner , at regeringens målsætning om at få 10.000 flere indvandrere i arbejde er lykkedes ? Ministeren . Det er rigtigt ,
ingen har et ambitiøst mål om , at beskæftigelsen for indvandrere og efterkommere skal øges med 10.000 personer frem mo
, at der er alt for mange med ikkevestlig baggrund , indvandrere og efterkommere , som rent faktisk står uden for arbe
det ganske kort : I 2008 havde vi 68.000 ikkevestlige indvandrere uden for arbejdsmarkedet , og det tal var i 2013 steg
V):Mener ministeren , at beskæftigelsesfrekvensen for indvandrere og flygtninge i den arbejdsdygtige alder er høj nok ?
: Mener ministeren , at beskæftigelsesfrekvensen for indvandrere og flygtninge i den arbejdsdygtige alder er høj nok ?
le . Tidligere har det været sådan , at flygtninge og indvandrere faktisk selv gik og sagde , at det var økonomisk attr
fortælle , at der i 2001 i Danmark var 14.111 ledige indvandrere / efterkommere fra ikkevestlige lande . Tallet stiger
nisteren har haft ansvaret for beskæftigelsen også af indvandrere / efterkommere . Der er jo ingen tvivl om , at alle s
, at man vælger at give ekstra penge til arbejdsløse indvandrere , som kan opnå 2.000 kr. ekstra skattefrit om måneden
tholde . Man har forhøjet ydelserne til udlændinge og indvandrere , man har lempet kravene for optjening af dagpenge ti
ennesker på pension og hæve ydelserne til arbejdsløse indvandrere . Derfor er mit spørgsmål til Socialdemokraternes ord
opbakning til ligestilling mellem kønnene , både hos indvandrere og hos efterkommere og hos etniske danskere . Man kun
om den måde , man ser på flygtninge på , og vi ser på indvandrere på , men her spørger jeg principielt set : Er det en
islamiske værdier , og hvor mange af de uintegrerbare indvandrere bor . Jeg formoder , at det gav bonus på valgnatten ,
fortæller , at muslimer er sådan og sådan , eller at indvandrere og efterkommere ikke har fortjent det samme som alle
ræddersyet , som jeg siger , til at ramme de såkaldte indvandrere . Det er jo sådan lidt blevet et begreb for alt mulig
-timersreglen og andre fattigdomsydelser , som holder indvandrere og alle andre fast i store problemer , og som forværr
r nylig sagde , at ud over nogle få grønthandlere har indvandrere ikke bidraget med noget . Altså , det betyder jo noge
ørselsindkomst , fordi det – og jeg citerer – går til indvandrere og efterkommere . Jeg skal bare lige have opklaret ,
f.eks. folkepensionister ? Der vil jo være masser af indvandrere , som bor i Danmark , som har knoklet i Danmark hele
sindkomst , fordi det også går til – og jeg citerer – indvandrere og efterkommere . Nu siger ordføreren , at det skal k
dtalelse om , at det også skulle være efterkommere og indvandrere . Det er trods alt godt at høre . Det er mærkeligt ,
r kvinder ? Ordføreren . I forhold til spørgsmålet om indvandrere og efterkommere er det jo typisk sådan , at udlænding
, men ellers i statistikkerne bliver registreret som indvandrere og efterkommere . Så er der nogle af dem , der har få
p 20 år . Jeg har arbejdet på skoler med rigtig mange indvandrere , og det er på deres vegne , jeg føler mig dybt kræn
Ordføreren siger : Jeg har været skolelærer for mange indvandrere . Indvandrere , altså drenge og piger , som er født o
er : Jeg har været skolelærer for mange indvandrere . Indvandrere , altså drenge og piger , som er født og opvokset i D

Gruppering af tekster
Computationel tekstanalyse illustreret



Gruppering af tekster
Computationel tekstanalyse illustreret



Dictionary metoder

- Prædefineret ordliste relateret til koncepter
- Genbrug af eksisterende ordlister/dictionaries (fx til sentiment-analyse)

Superviseret maskinlæring

- Kræver at tekster (hele tekster, afsnit, sætninger) er kategoriseret i forvejen (hvilket koncept afspejler teksten?)

Involverer en eller anden form for *usuperviseret maskinlæring*.

Gruppering på tekstniveau

- Klyngeanalyser
- Topic models

Gruppering på ordniveau

- Tekst som vektorer (fx word embeddings)

Brug af eksisterende sprogmodeller (specifikt dependency parsing og evt. named entity recognition) til at udlede sammenhænge i tekster.

Subjekt -> handling -> objekt ("Subject-action-object")

- "Faktiske"handlinger
  - Fra fx nyhedsmedier
- Formodede handlinger
  - Italesættelse af befolkningsgrupper

danskerne	savner	kultur
danskerne	turde	søge
danskere	føler	glæde
danskere	føler	mistrives
danskerne	se	er
dansker	få	den
danskerne	have	lov
dansker	får	penge
danskere	får	penge
dansker	får	mere
danskerne	gå	vej
danskere	får	adgang
danskere	spare	titusinder
danskerne	gå	alderdom
danskere	gå	alderdom
danskere	valgte	købe
danskerne	fandt	trøst
danskerne	vist	sig
danskere	genkende	billede
danskere	have	nyt
danskere	liggende	som
danskerne	betale	skat
danskere	får	adgang
danskerne	får	feriepenge
danskerne	se	forskel
danskerne	se	ramte
danskerne	støtter	ene
danskerne	have	feriepenge
danskerne	tage	ejerskab
danskerne	får	lov
danskere	betale	topskat
danskerne	hæve	feriepengene
danskerne	mister	arbejde

- Behandling og bearbejdning af “naturligt sprog” ved hjælp af computerteknologi
- Teknikker der involverer statistiske metoder til at forstå tekst - med eller uden lingvistiske indsigter
- Et sæt af værktøjer der tillader, at computere kan bearbejde og udlede information fra lingvistisk data (tekst eller tale)
- Krydsfelt mellem datalogi og lingvistik
- Involverer i stigende grad brug af maskinlæringsteknologi

Eksempler fra hverdagen: ChatGPT, Tale-til-tekst applikationer, tekstforslag i beskeder, autokorrektur, oversættelsestjenester (Google Translate)

Opdeling af tekst til enkelte “tokens” (ordenheder)

```
‘Politiet har givet borgerne råd‘ -> [‘Politiet‘, ‘har‘, ‘givet‘,
‘borgerne‘, ‘råd‘]
```

- Evt. frasortering af tegnsætning
- Evt. konvertering til små bogstaver (lower-casing)

Udfordringer:

- Egenavne i flere ord: Høje Taastrup
- Sammensatte ord med bindestreger

Typisk referer tokenization til at opdele tekster i enkeltord.

Afhængig af formål, kan det være relevant at se på tokens som "ordpar" (bigrams) eller "ordsamlinger" (n-grams).

Bigram eksempel:

`‘Politiet har givet borgerne råd‘ -> [(‘Politiet‘, ‘har‘), (‘har‘, ‘givet‘), (‘givet‘, ‘borgerne‘), (‘borgerne‘, ‘råd‘)]`

Typisk vil man ikke behandle ord med forskellig gradbøjning som forskellige ord

Simpleste måde at ensrette: omdanne til ordstammen

- køre -> kør
- kører -> kør
- køren -> kør
- kørte -> kørt

Flere algoritmer til at lave stemming - *dog ikke altid enige!*

- Snowball algoritme (sprogfølsom): køren -> kør
- Porter algoritme (ikke sprogfølsom): køren -> køren

Konverterer ordet til grammtiske stamme (alternativt til stemming)

- køre -> køre
- kører -> køre
- køren -> køren
- kørte -> køre

Lemmatizere er af nødvendighed sprogafhængige

Findes typisk som “dictionary-modeller” (nogen har lavet opgørelser af, hvilke ord har samme grammatiske stamme)

Mange ord indgår i tekst uanset hvad teksten handler om (bindeord, stedord, forholdsord o.l.)

Refereres til som “stopord” - frasorteres da de typisk ikke giver indblik i, hvad tekster handler om

Stopord kan både være generelle (bindeord, stedord, forholdsord) eller kontekstspecifikke (fx ‘ordfører’ i referater fra Folketinget).

Opdeler ord i ordklasser: navneord, udsagnsord, tillægsord osv.

Brugbart til at udlede meningsfulde ord i tekst (tillægsord siger typisk mere om teksten end forholdsord).

Er nødt til at være sprogafhængige *og* kontekstafhængige (fleste er baseret på maskinlæring i dag).

`‘Politiet har givet borgerne råd‘ -> ‘Politiet_NOUN har_AUX givet_VERB borgerne_NOUN råd_NOUN‘`

Dependency parsing udleder sætningskonstruktion: grundled, udsagnsled, genstandsled.

Hjælper til at reducere tvetydighed i tekst - “hvem gjorde hvad til hvem?”

Brugbart til at udlede meningsfulde dele af teksten (fx fokusere på visse ord, der indgår som genstandsled i sætninger).

Er nødt til at være sprogafhængige *og* kontekstafhængige (fleste er baseret på maskinlæring i dag).

`‘Politiet har givet borgerne råd‘ -> ‘Politiet_nsubj har_aux givet_ROOT borgerne_obj råd_obj‘`

Introduktion til sprogmodeller

En sprogmodel (“language model”) er kort sagt en model til at forudsige ord. Der er tale om modeller, der er trænet til at forstå forskellige sprog, og kan derfor både bruges til at producere tekst på et bestemt sprog eller analysere opbygningen af en tekst.

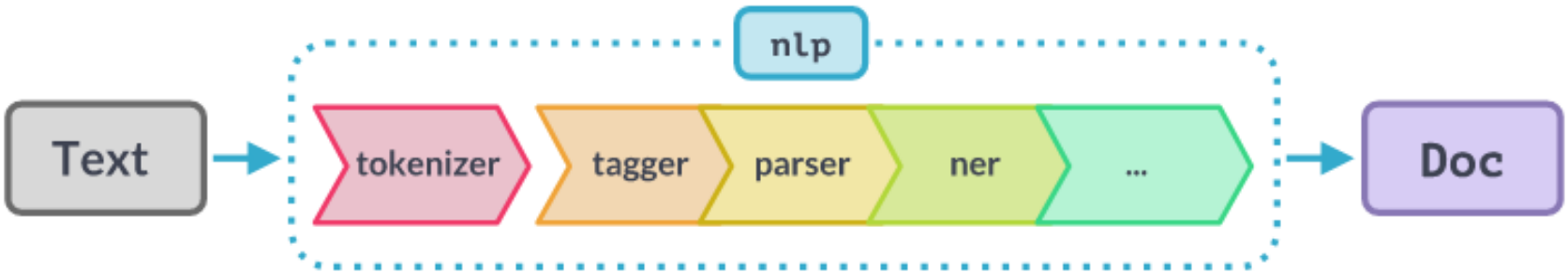
Lidt forsimplet er en sprogmodel en model, som er trænet til at genkende ordtyper, entiteter og sætningskonstruktion, som derved kan prædiktere disse informationer i tekst (for det meste inden for samme sprog).

"spaCy" indeholder forskellige sprogmodeller - herunder en dansk sprogmodel.

Overordnet virker spaCy ved, at man specificerer en sprogmodel samt nogen "processors", som modellen skal indeholde.

SpaCy's sprogmodeller indeholder blandt andet:

- Tokenizer (inddeling i enkeltord)
- Lemmatizer (konvertering til navneform)
- Part-Of-Speech tagging (POS-tagging) (identificering af ordtyper)
- Dependency parsing (sætningskonstruktion)
- Named-Entity-Recognition (NER) (udledning af "named entities", fx personer og organisationer)



- Et spaCy pipeline tager én tekst ind og giver et doc-objekt tilbage
- NLP-komponenter kan tilføjes og fjernes (obs på afhængigheder)
- spaCy pipelines er tilgængelige gennem eksisterende sprogmodeller

1. Indlæs sprogmodel
2. Analysér tekstykke
3. Inspicér resultater

Når modellen anvendes på et stykke tekst, kan de forskellige værdier og information, som er udledt af teksten, tilgås som attributes (et attribute for token, et for lemma, et for part-of-speech-tag osv.).

Brug af spaCy i Python (live-coding)

AALBORG UNIVERSITY
DENMARK

Brug af dictionary metoder og sentiment analyse i Python (live-coding)

AALBORG UNIVERSITY
DENMARK

Teori i computationel tekstanalyse

Access to greater volumes of data—and more complex data, such as text (as well as maps, images, and audio)—has necessitated more theory, not less (Bonikowski & Nelson, 2022)

Teori i computationel tekstanalyse

- Udvælgelse af corpus
- Valg i pre-processing
- Udvalg af ord til at måle koncepter (i dictionaries)
- Manuel kategorisering af tekster
- Fortolkning af tekstgrupperinger og ordforbindelser
- Validering

Corpus som en stikprøve

- Corpus som en bestemt "genre-> Kræver domæneviden!
- Hvad og hvem repræsenterer data?
- Computationel tekstanalyse involverer ofte genanvendelse -> Data skabt til andet formål end forskning

Indsamling af relevant data

- For eksisterende samlinger og arkiver: Hvem har samlet det? Hvad repræsenterer det?
- For sociale medier: Hvad er tilladt? Hvad kan man få adgang til? Hvad mangler? Hvem "taler"?
  - Selektionsproblemer
  - Mangel på baggrundsoplysninger

Match mellem forskningsspørgsmål og metode

- Flere modeller og metoder kan opnå mere eller mindre det samme teknisk set
- Metodens brugbarhed skal evalueres ift. problemstillingen og den teoretiske indsigt, som man ønsker at opnå med metoden
- Problemstillingen skal derfor altid være styrende for metodevalget (ikke omvendt)

Kombination af metoder og teknikker

- Ofte kræver computationel tekstanalyse en kombination af flere metoder og teknikker
  - Fx brug af sprogmodel til pre-processing -> inddeling af tekster i grupper -> test af sammenhæng mellem baggrundsvariable for tekster og gruppering
- Formålet med kombination af metoder er ikke at belyse data fra flere vinkler, men at sammensætte meningsfulde workflows

  [T]urning text into data involves a cascade of interlocking path-dependent choices and so it is rare that particular choices will be right or wrong without the context of choices made before and after. (Evans, 2022)
- Ofte iterativ og abduktiv proces
  - Deduktivt: Udlede specifikt koncept i corpus (dictionaries, superviseret maskinlæring)
  - Induktivt: I hvilke kontekster optræder det specifikke koncept i corpus (usuperviseret maskinlæring, topic modelling)

Brug af eksisterende modeller og ressourcer

- Tilgængelighed af sprogressourcer (corpora, modeller, dictionaries) skaber bedre muligheder for computationel tekstanalyse.
- Eksisterende sprogressourcer er dog altid lavet af *nogen* med *noget*.
  - Hvem har lavet det?
  - Hvad er deres kvalifikationer?
  - Hvor godt er det beskrevet?
  - Hvilke data og ressourcer er blevet brugt? Hvor stammer de fra?
- Selv de bedste sprogressourcer er afhængige af, at andre har lagt meget arbejde i manuel annotering

Validering i computationel tekstanalyse et spørgsmål om *gennemskuelighed*

- Valg af corpus og hvad det repræsenterer
- Valg i databehandling og hvorfor
- Valg af metode og hvad det skal indfange
- Hvordan resultater er valideret

Principper for automatiseret tekstanalyse (Grimmer & Stewart, 2013)

- Quantitative methods augment humans, not replace them
- There is no globally best method for automated text analysis
- Validate, validate, validate

I computationel tekstanalyse er det ofte tilfældet, at både data og metoder er taget fra andre discipliner eller ikke fra en forskningskontekst - kræver omhu, validering og kobling til samfundsvidenskabelig teori!

Litteratur

Bonikowski, B., & Nelson, L. K. (2022). From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research. *Sociological Methods & Research*, *51*(4), 1469–1483. doi: 10.1177/00491241221123088

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, *2*(2), 205395171560290 8. doi: 10.1177/2053951715602908

Evans, J. (2022). From Text Signals to Simulations: A Review and Complement to Text as Data by Grimmer, Roberts & Stewart (PUP 2022). *Sociological Methods & Research*, *51*(4), 1868–1885. doi: 10.1177/00491241221123086

Franzosi, R. (1989). From Words to Numbers: A Generalized and Linguistics-Based Coding Procedure for Collecting Textual Data. *Sociological Methodology*, *19*, 263–298. doi: 10.2307/270955

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. doi: 10.1093/pan/mps028

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139524773

Macanovic, A. (2022). Text mining for social science – The state and the future of computational text analysis in sociology. *Social Science Research*, *108*, 102784. doi: 10.1016/j.ssresearch.2022.102784

Stuhler, O. (2022). Who Does What to Whom? Making Text Parsers Work for Sociological Inquiry. *Sociological Methods & Research*, *51*(4), 1580–1633. doi: 10.1177/00491241221099551