



AUL ARTS Nobelparken

DET KGL.
BIBLIOTEK

Royal Danish Library

Introduktion til digitale metoder for litteraturstudier - text mining og distant reading (nogle) ressourcer, værktøjer og tilgange

Karoline Liv Vildlyng, informationsspecialist

Hej!

- Karoline Liv Vildlyng
- Informationsspecialist på Det Kgl. Bibliotek i afdelingen for Biblioteksservices og Partnerskaber
- Cand. Mag. i moderne kultur og kulturformidling, specialiseret i digital kultur og digitale metoder
- Jeg underviser og arbejder også som kontaktbibliotekar for filosofi og idehistorie

Hi!

1 diskussionsøvelse

Find sammen i par af to og diskuter dine forventninger til workshoppen

1. Hvad forventer du at vide efter dagens workshop som du ikke allerede ved?
2. Hvad er dine erfaringer med at arbejde med digitale metoder?



Hvad ved du efter dagens workshop?

Du ved hvad Voyant er og hvordan du kan bruge de primære funktioner i programmet

Du ved også:

- Hvad en wordcloud er og hvordan en netværksgraf ser ud
- (mere om) hvordan en computer læser
- Hvordan du kan bruge stopord til at rense dine data
- Hvordan du laver og analyserer et corpus
- Hvordan du analyserer, kontrasterer og sammenligner store mængder af tekst gennem computational text analysis aka *text mining*



Kompetencemål for studerende

At du:

- Får grundlæggende kendskab til text mining som koncept og metode, samt hvorfor du måske gerne vil inkorporere digitale metoder i din akademiske praksis
- Udvikler kritisk bevidsthed om Voyants interface, værktøjer og funktioner, samt dets værdi som metodeværktøj
- Føler dig tryg ved at udforske digitale metoder videre på egen hånd

Program for dagens workshop

Første blok - intro (45 minutter)

- Intro til natural language processing og text mining
- Centrale forståelser omkring digitale metoder for litterære studier og litteraturhistorie
- Hvorfor *mixed methods*? Eventyr-eksempel

Anden blok - hands-on code along (45 minutter)

- Github
- Demonstration af Voyant
- Gruppeøvelse: hands-on text mining med *Frankenstein* (1818 og 1831)

Tredje blok - eventyr, kritik og diskussion (45 minutter)

- Gruppeøvelse: Kontraster og sammenlign to corpora af skønlitteratur
- Gruppediskussion: hvad kan digitale metoder, hvad er begrænsningerne og hvordan er de relevante for jer?

En intro til text mining og fjernlæsning

Aka 'beregningsbaseret' tekstanalyse:

- Et andet term for text mining eller *natural language processing*
 - En proces hvor information udtrækkes fra (store mængder) tekst som fx romaner, monografer, artikler, hjemmesider etc. aka et *corpus*

En intro til text mining og fjernlæsning

- Generelt indvolverer text mining at identificere mønstre ved at udregne ordfrekvenser eller associerende links mellem ord, sammenfald og ordtæthed
- Er en såkaldt mixed methods approach der kombinerer kvalitative og kvantitative tilgange til forskning inden for humaniora
- Det er ikke så nyt som vi tror



En intro til text mining og fjernlæsning

Fjernlæsning var formaliseret som sådan og introduceret til felten af Franco Moretti i 2005.

Men kvantitativ tekstanalyse som metodepraksis har været udøvet i årtier. Når papirbaserede værker er mediummet, er det en voldstomt tidskrævende praksis.

Et eksempel - Ruth Bottigheimer i 1986:

- Analyserede tale vs stilhed på tværs af eventyr, herunder Brd. Grimm
- Optalte ordfrekvenser relateret til tale, samt hvilke mønstre de optræder med i relation til karakterens køn og rolle



alamy

Image ID: ECXY4
www.alamy.com

#1 Spiløvelse - nøgleordsbingo

1. Skriv ned hvad du tænker hvert term er/betyder

2. Derefter gennemgår vi dem sammen

3. Tæl hvor mange rigtige (eller næsten rigtige) du har undervejs

Corpus

OCR

Ordforrådstæthed

Natural language processing

Token

Stopord

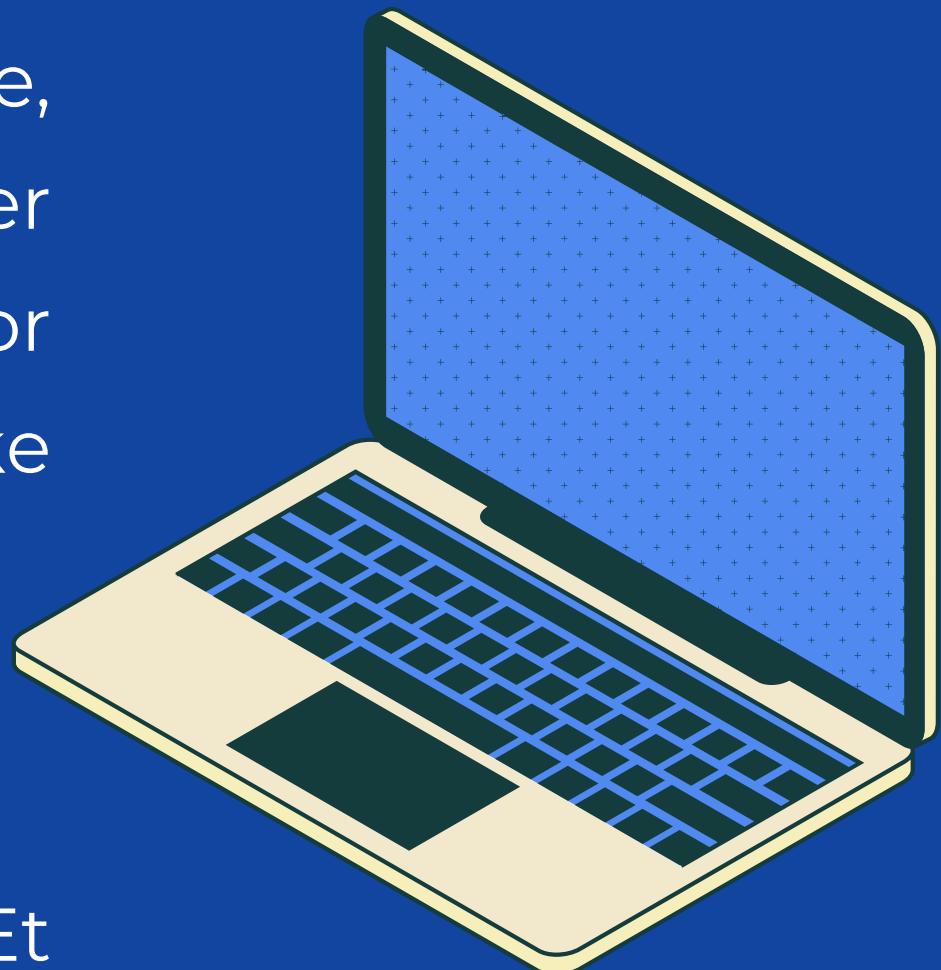
Nøgleord

Corpus; flertal. Corpora:

En større, og struktureret mængde af tekster. Kan være i flere eller bare ét sprog. Er ikke sammenligneligt med en database, antologi eller arkiv. Kan være et samlet forfatterskab eller værker centreret om samme emne. Godt format for hypotesetestning eller statistisk analyse af teksters lingvistiske strukturer, frekvenser etc.

Natural language processing:

Et 'naturligt' sprog er engelsk. Eller spansk eller dansk. Et naturligt sprog er ikke Javascript eller Python eller et Excel-ark som en computer kan læse og afkode.



Nøgleord

Token:

Er hvad vi kalder et 'ord' inden for text mining.

Det er alt det, der ikke er et mellemrum.

Det er en **enhed/unit**.

For at computere kan afkode tekst og oversætte corpus til tal, skal teksten formaliseres i enheder.

Kan sammenlignes med at skære en tekst op i bidder af symboler, ord og tal.



Nøgleord

OCR (optical character recognition):

Brug af computerteknologi til at konvertere scannede billeder af maskinskrevet, printet eller håndskrevet tekst til maskinelt aflæselig tekst. OCR-fejllæsninger er også grunden til at du måske kan finde fejl i dine datasæt.

Ordforrådstæthed:

Et mål for ordforråd i sammenligning med længden af teksten. Sammenhold af antal ord i dokumentet med antal unikke ord i dokumentet.

En lavere tæthed i ordforråd indikerer kompleks tekst med mange unikke ord og en højere indikerer mere enkel tekst med mange repetitioner af ord.

Nøgleord

Stopord:

Ord der gentages med høj frekvens uden at tilføje mening alene fordi de er gentaget.

Og, eller, i, når, hen.

'Fyldeord' der tager plads, men ikke giver analytisk indsigt.

Når vi analyserer et corpus skal vi være sikre på at have den mest optimale **stopordsliste**.

Optimal, i denne kontekst, betyder den mest optimale for jeres undersøgelse.



Hvorfor mixed methods?

- Dagens eksempel-corpora er eventyr af H. C Andersen og Brdr. Grimm
- Et fælles kendetegn ved eventyr er hvordan de præsenteres. Ordene selv fortæller narrativet med begrænset indre psykologi beskrevet fra karakterenes side
- Dette gør eventyr velegnede for kvantitativ tekstanalyse. Den narrative struktur er enkel og læseren tilføjer subteksten



Hvorfor mixed methods?

- Mange forskningsprojekter i kvantitativ tekstanalyse af eventyr har været inden for kønsstudier

Et eksempel er Jeana Jorgensen og Scott Weingart (2013):

- Gennem digitale, kvantitative værktøjer, har de analyseret alder og køn i eventyr i relation til hvordan karakterenes fysiske fremtoning beskrives
- De fandt at kvindelige karakterer ofte bliver beskrevet mere end de mandlige og at ældre karakterer beskrives ud fra fysik mere end de yngre -> det narrative omdrejningspunkt er mandligt og ung
- Jorgensen (2014) fandt at beskrivelserne af kvinder prioriterer skønhed, moral, blod, hår og hud hvor beskrivelser af mænd involverer vold, transformation, størrelse og alder

Hvorfor mixed methods?

- Eventyr er delvist defineret ved at blive fortalt og genfortalt i utallige versioner og de eksisterer i mange udgaver over tid. Dette gør dem også velegnede for kvantitativ tekstanalyse
- Eksempel på metodedesign: Brug Voyant til at analysere forskellige versioner af samme eventyr over tid. Hvordan ændres ordfrekvensen gennem fire forskellige fortællinger af Rødhætte?
- Dette kan også gøres med en version af det originale eventyr vs en ny genfortælling af samme historie. Rapunzel og Tangled eller Snedronning og Frozen. Hvordan ville ordforrådstætheden skifte eller ordfrekevenser rykke sig?

Hvorfor vi ikke analyserer *Frankenstein* (1931) i dag...

- Som udgangspunkt er værker i Danmark beskyttet af lov om ophavsret frem til 70 år efter forfatterens død
- Hvis et værk er omfattet af ophavsret må du kun bruge det til private formål, ikke til distribution
- Så: Vi (på Det Kgl. Bibliotek) bruger The Gutenberg Project til at tilgå værker der ikke er under ophavsret



Pause - 15 minutter



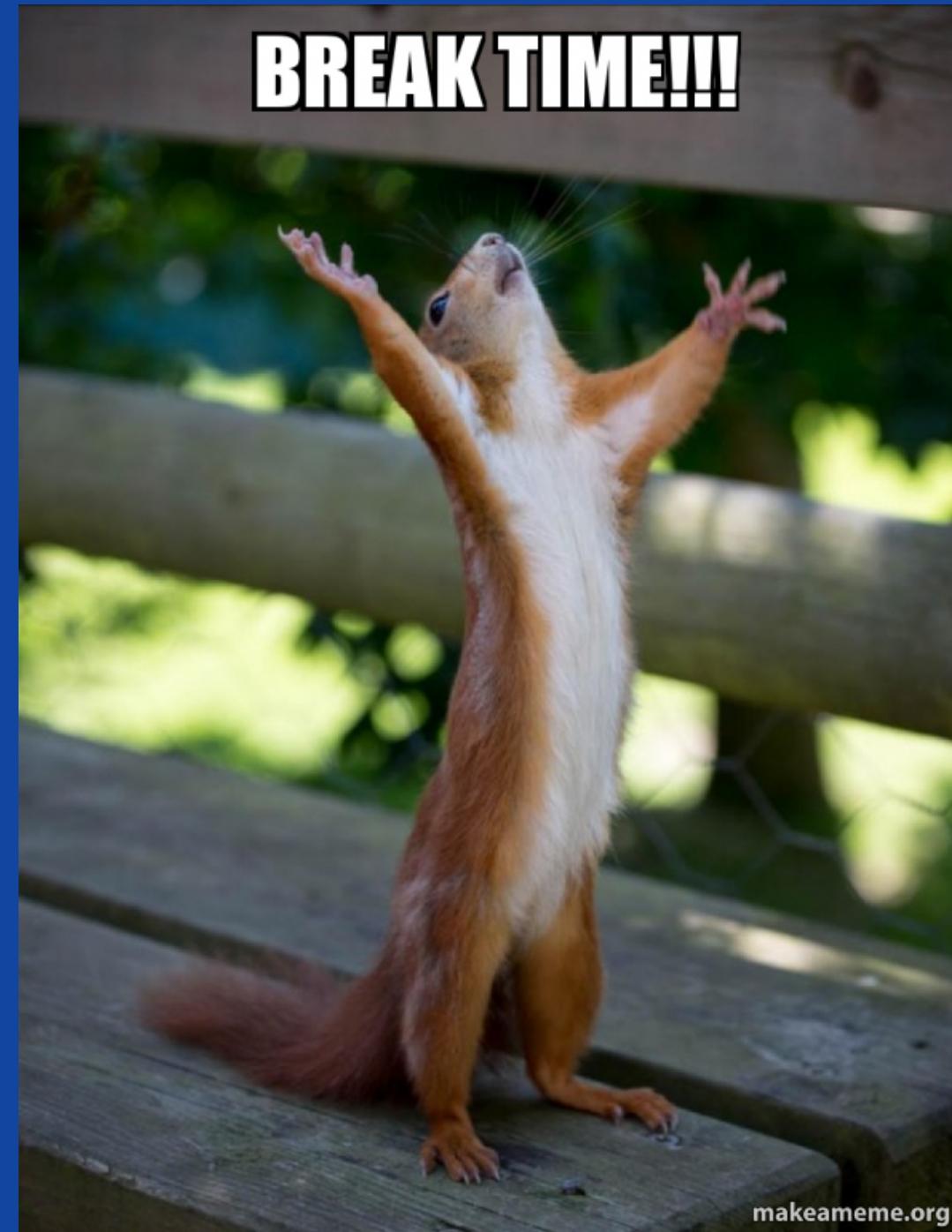
Voyant gennemgang og hands-on text mining

Første step

Gå til:

[https://github.com/karolinevildlyng/
Introduction-to-Voyant](https://github.com/karolinevildlyng/Introduction-to-Voyant)

Pause - 15 minutter



Kritik og refleksion

- Informations- og kommunikationsteknologier har fundamentalt ændret hvordan vi lærer - og læser?
- Moretti (2005) beskriver fjernlæsning som '*a little pact with the devil: we know how to read texts, now let's learn how not to read them*'.



#2 diskussionsøvelse - 5 minutter

Find sammen i par af to

1. Diskuter: Er det at uploadet corpus til Voyant og analysere tal, grafer og oversigter, *at læse*?
2. Hvis ja, hvorfor? Hvis nej, hvorfor ikke?

Kritik og refleksion

"Literary criticism is, in many ways, the art of telling stories about stories. Using textual evidence, convincing critics weave alternative narratives around texts that re-contextualize them forever, making it impossible to re-read a work in the same way again. Thus, criticism is a performance of a new narrative and a deformation of the original, now-reshaped text. Literary criticism routinely undertakes, then, what Lisa Samuels and Jerome McGann call 'deformance'."

Martin Paul Eve (2022) *The Digital Humanities and Literary Studies*

Kritik og refleksion

- De æstetiske og kvalitative videnskaber kritiseres ofte for at være 'anekdotiske'

"The literary scholar of the twenty-first century can no longer be content with anecdotal evidence, with random “things” gathered from a few, even “representative,” texts. We must strive to understand these things we find interesting in the context of everything else, including a mass of possibly “uninteresting” texts."

Matthew L. Jockers (2013) *Macroanalysis*

- Men kvantitative data er heller ikke neutrale...

Kritik og refleksion

Ingen software er perfekt og en computer kan ikke læse som et menneske kan (endnu)

- Eksempel: Måden hvorpå Voyant udregner sætningslængde skal ses som en tilnærrelse fordi det er svært at skelne mellem et punktum og et komma eller semikolon
- Stanley Fish (2012) pointerer at hovedproblemet med fjernlæsning er at "*first you run the numbers, and then you see if they prompt an interpretive hypothesis. The method, if it can be called that, is dictated by the capability of the tool*"
- En kritik er at værktøjet kan ende med at dikttere *metoden* - istedet for omvendt

Kritik og refleksion - og svar?

Digitale metoder erstatter ikke traditionelle metodologier i de humanistiske videnskaber

- De er et supplement til eksisterende videnskabelige forskningsprakssiser
- Gennem dem kan vi skabe indsigt vi ellers ville have svært ved at få
- De kan være en vej til at automatisere ellers manuelle arbejdsgange
- De kan være med til at skabe nye diskussioner og reflektioner over ellers velkendt materiale
- De kan være fejlbehæftede og have blinde vinkler helt ligesom traditionelle metoder

Tak for at lytte og deltage!

Biblioteket er her for at hjælpe jer - I er altid velkomne til at skrive til mig på

kavk@kb.dk
