DBC
D1G1TAL

AI som redskab i metadataproduktion

Eksempler fra DBC

Philine Zeinert og Noah Torp-Smith

Agenda

- ChatGPT og Sprogmodeller
- Metadata og vores AI-services
- Appeldata (buggi, læsekompas)
- Auto-index
    - Ontologi og auto-onto service
    - Tekniske overvejelser for en AI-indekseringsmodel
    - Demo af foreløbige modeller
    - Status quo
- AI eksperimenter

DBC
D1G1TAL

ChatGPT

- Vi kender alle til ChatGPT – f.eks. fra virtuelt kaffemøde i september
  - Det er **komplekst**
  - Vi skal **omfavne** men være **konstruktivt kritiske**
  - Eksperimentere og gentænke
- Pointer fra Jamboards:
  - Mangler **retningslinjer** fra enten moderinstitutioner eller endnu højere oppe.
  - **Kildekritik** – helt generelt
  - Hvordan skal vi **styre** de studerendes brug? Ansvar?
  - Skal der ændres "**workflow**" for forskning? Undervisning? Vejledning?

Sprogmodeller er andet end ChatGPT

- Klassifikation af en tekst
  - "Sentiment analysis" det klassiske eksempel fra undervisning i sprogmodeller.
  - Automatisk udledning af hvilken genre en tekst er - krimi/kærlighedsroman/videnskabelig artikel?
  - (hvad "handler" en tekst om?)
- NER: identifikation af "vigtige bestanddele" af en tekst
  - Genkende personer/steder/tidsangivelser [DEMO]
  - Udfordring: det virker bedst på engelsk
  - Fintuning, så vi også kan identificere andre ting.
- Find materialer der "minder om"
  - Recommendere
  - Relaterede emneord/søgninger

DBC
D1G1TAL

NER Screenshot (hvis demo ikke fungerer)

[65]:
```
import spacy
from spacy import displacy
nlp = spacy.load("en_core_web_sm")
#doc = nlp("He works at Google.")
```

[84]:
```
text = """When Sebastian Thrun started working on English self-driving cars at Google in 2007, few people outside of the company took him seriously.
“I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun,
now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode earlier this week. The interview took place in German because of Thrun's past.
A little less than a decade later, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand,
to secure their place in the fast-moving world of fully automated transportation."""

#text = """In ancient Rome, some neighbors live in three adjacent houses. In the center is the house of Senex, who lives there with wife Domina, son Hero, and several slaves, includin

#text = "He works at Google"
```

[85]:
```
doc = nlp(text)
```

[89]:
```
sentence_spans = list(doc.sents)
displacy.render(sentence_spans, style="ent", options={"ents":["PERSON", "ORG", "NORP", "DATE", "NOUN", "LANGUAGE", "VERB"]})
#displacy.render(sentence_spans)
```

When Sebastian Thrun PERSON started working on English LANGUAGE self-driving cars at Google ORG in 2007 DATE, few people outside of the company took him seriously.

“I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun PERSON,

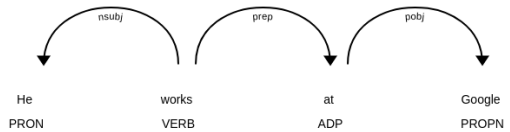
now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE.

The interview took place in German NORP because of Thrun's past.

A little less than a decade later DATE, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand,

to secure their place in the fast-moving world of fully automated transportation.

[90]:
```
text = "He works at Google"
doc = nlp(text)
displacy.render(doc)
```



[ ]:

DBC D1G1TAL

DBC
D1G1TAL

Brug af metadata i vores AI-services

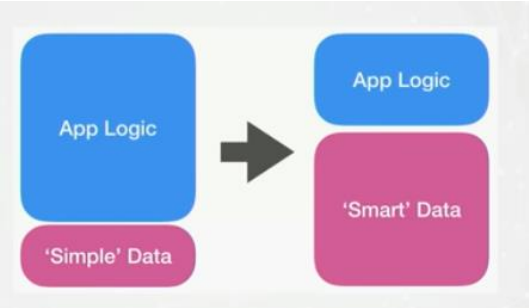
Metadata er over det hele!

- Metadata bruges til **søgning**
  - Folk søger på forfattere/emner/titler
- Metadata bruges til **suggestions**
  - Vi skal (hurtigt!) foreslå forfattere/emner/titler
- Metadata bruges til **recommendations**
  - Hvad betyder det at noget "minder om" - information fra metadata!
- Metadata bruges til at **generere/foreslå** mere metadata
  - Machine Learning/AI har brug for *træningsdata*

Betydning af metadata

Kvalitet af metadata er vigtigt!

- Der kan være deciderede fejl eller uhensigtsmæssigheder i data
- Data kan være struktureret på en måde vi ikke har forudset
- Der kan være data, vi ikke har valgt at inkludere i søgemaskinen



DBC
D1G1TAL

Perspektiv: natursprogssøgning

- Eksempel: - gøre brug af stemningsord til fx understøttelse af "Nye spændende bøger"

Vision: Bruge NER til at analysere søgning

- Nye franske kogebøger
- Norske krimier
- Dansk film
- Jussi lydbog
- Tove Ditlevsen
- Film 2023

Hvis vi hurtigt kan identificere sprog/materialetyper/personer etc, så kan vi give bedre søgeresultater.

DBC
D1G1TAL

Appeldata

Brug af stemningsdata i vores AI-Services

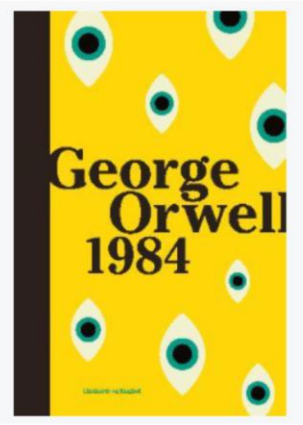
George Orwell - 1984

Author: George Orwell

Title: 1984

Genre/form:
dystopia, stories about the future, novel

Subjects:
the future, civil rights, surveillance society, the state



Mood	Fictive characters	Language	Structure	Pacing
• upsetting • gloomy • philosophical • thought-provoking	• Winston Smith • Big Brother	• plain language	• 3rd person narrator	• lingering

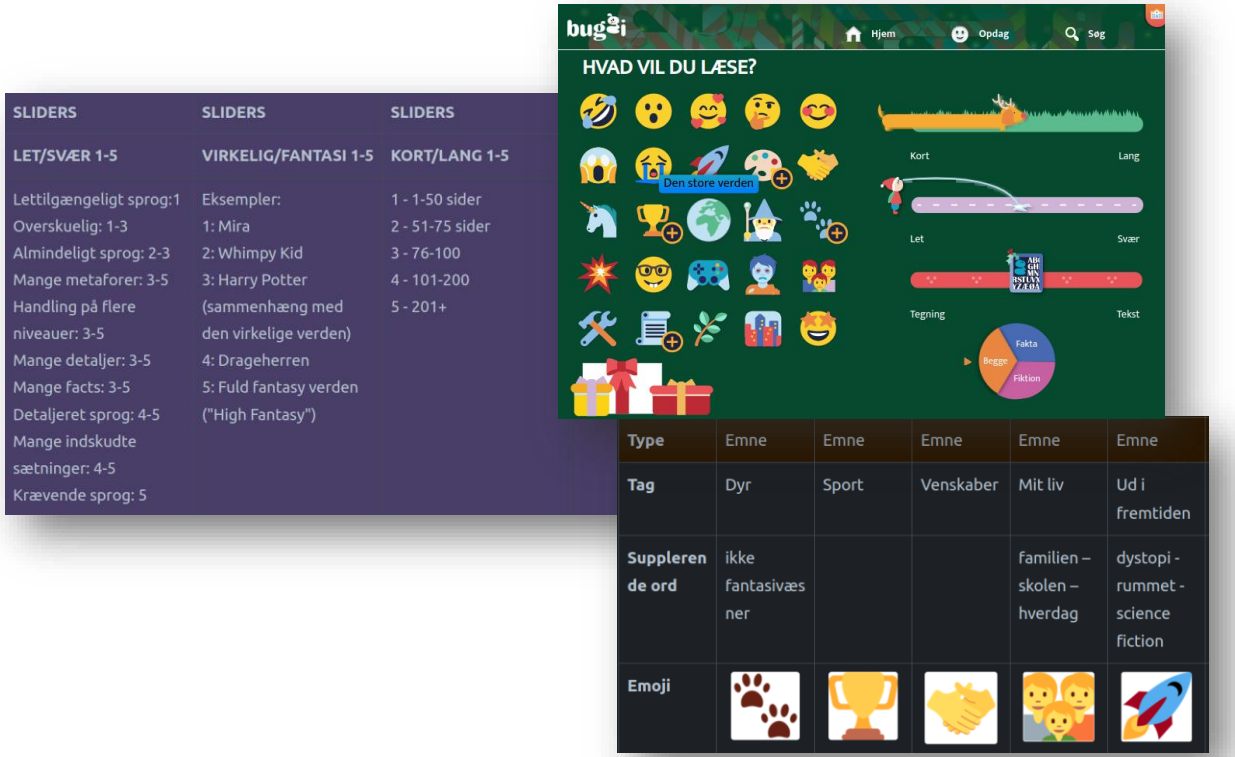
DBC
D1G1TAL

Appeldata Børn

Appeldata til børn bliver en del af de bibliografiske poster med egen felt i danMARC2.

Appeldata for børn består af:

- længde (kort/lang)
- sværhedsgrad (let/svær)
- mængde af illustrationer (tekst/tegninger)
- univers (virkelig/fantasi)
- stemningsord
- emne
- genre



DBC D1G1TAL

Buggi

Brug af metadata i buggi:

- Søgemaskine
- Suggesterfunktion

- **Anbefalinger**

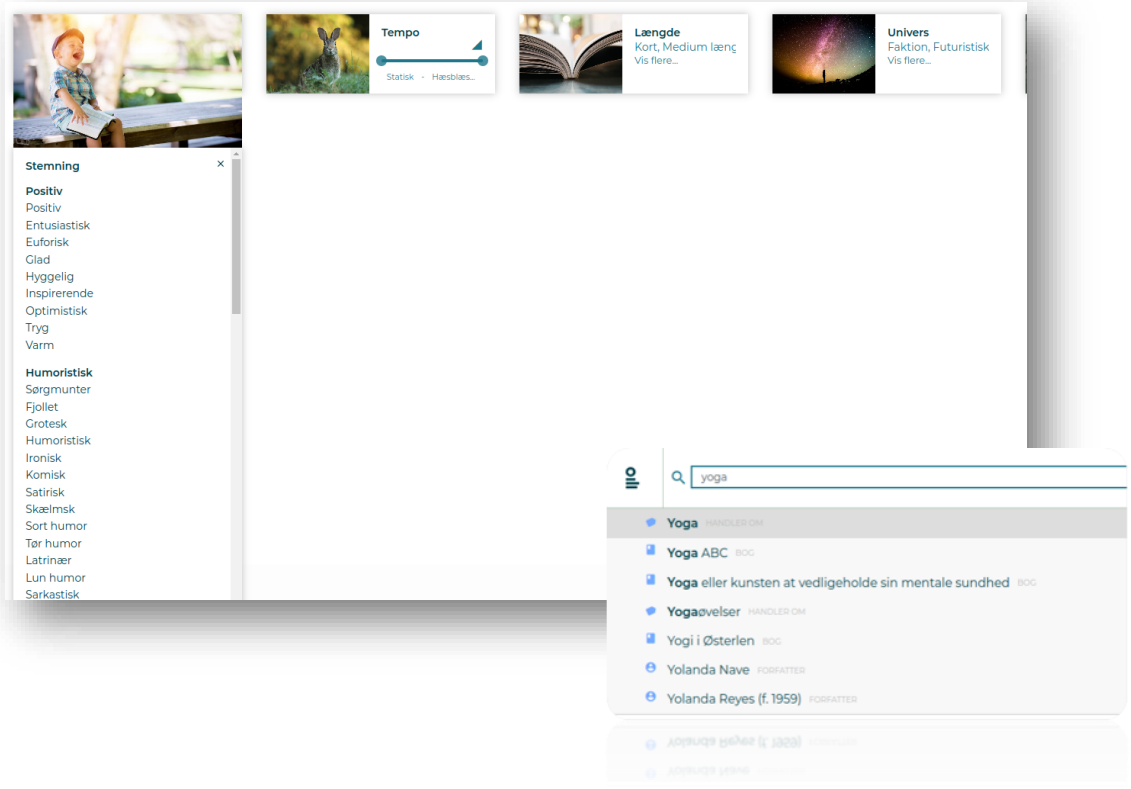
https://buggi.dk/school

Appeldata Voksne

Læsekompas
LAD BIBLIOTEKET INSPIRERE DIG

Voksne stemningsdata består af:

- skrivestil og struktur
- fortællerstemme
- længde
- tempo
- stemning inden for 9 overkategorier
- univers og miljø
- ...



DBC
D1G1TAL

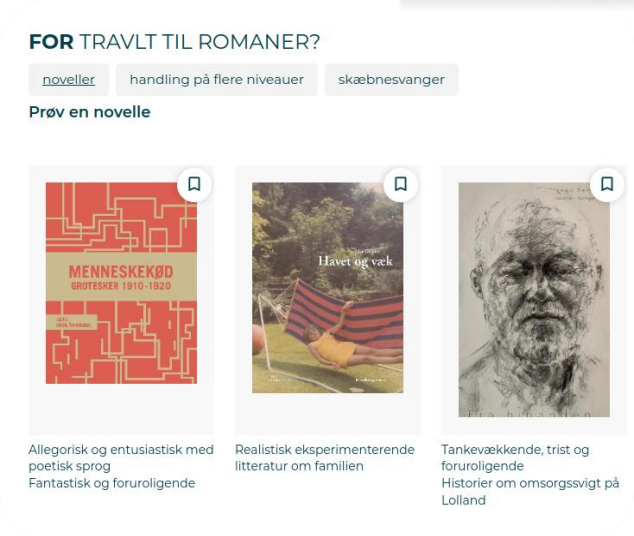
Læsekompas

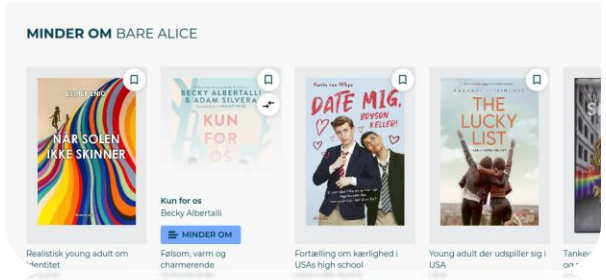
Brug af metadata i læsekompas:

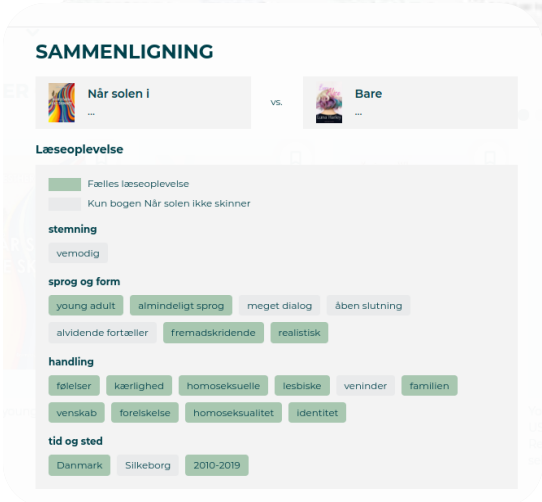
- Søgemaskine
- Suggester
- **Anbefalinger baseret på tags**
- **Anbefalinger baseret på titler**

(Minder om)

https://laesekompas.dk/







DBC
D1G1TAL

DBC
D1G1TAL

Auto-Index

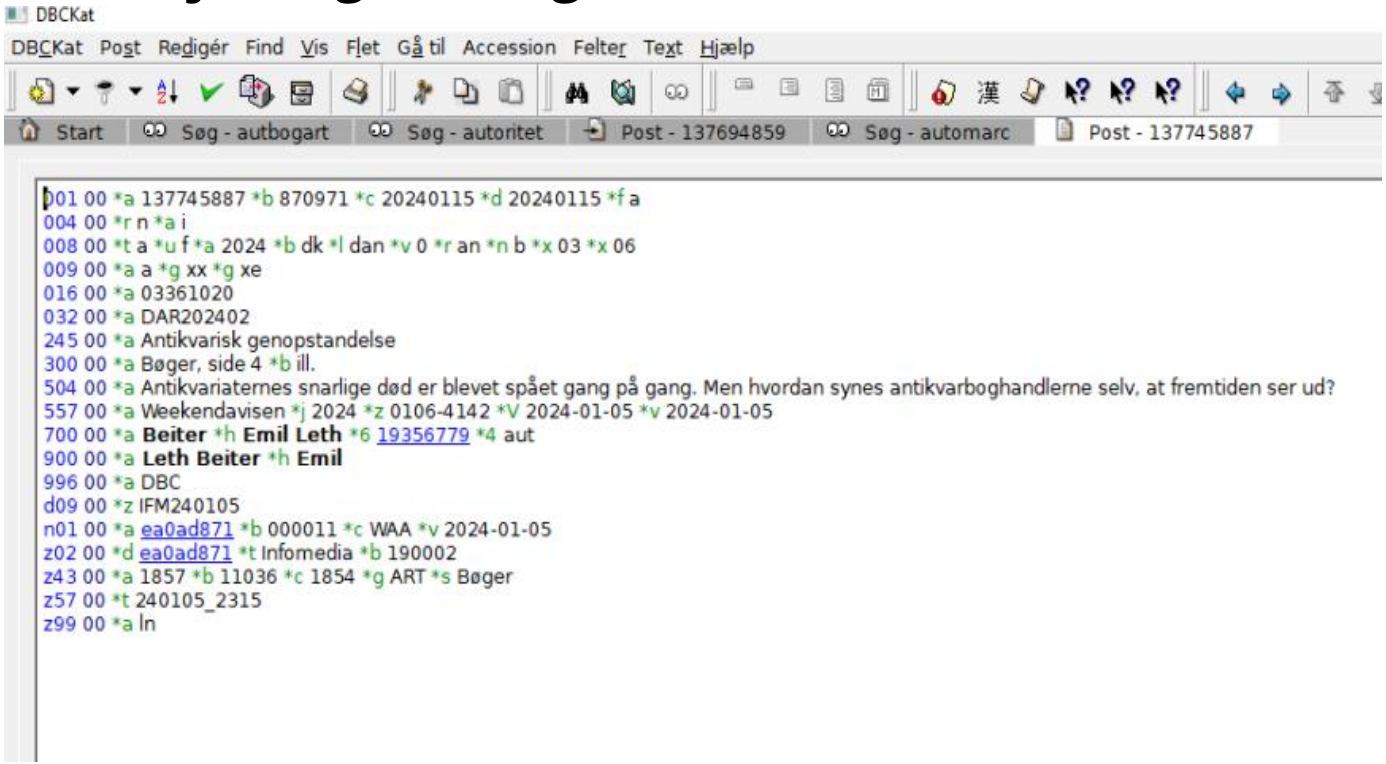
Et AI-værktøj til produktion af metadata

Auto-Index: Kontekst

- Dansk Artikelindex
- Cirka 29000 artikler om året, så i størrelsesorden 150 om dagen.
- Data kommer fra InfoMedia i fuldtekster.
- Der sidder "rigtige mennesker" og laver klassemærker, emner og andre ting for artiklerne.
- Vi har en mængde af "kontrollerede emneord" som man "helst" skal holde sig indenfor.

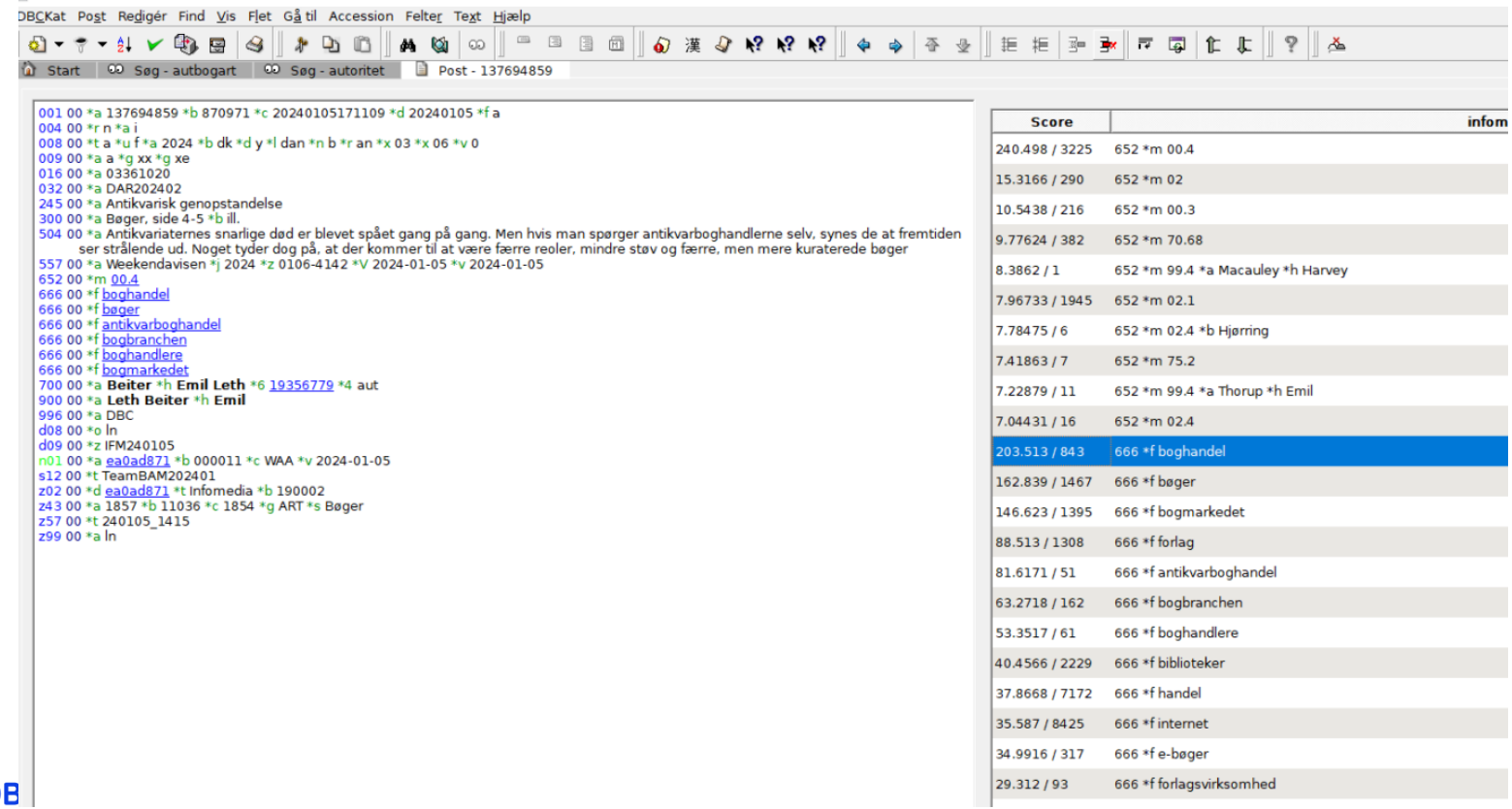
DBCKat/FBIKat

Værktøj til registrering af metadata



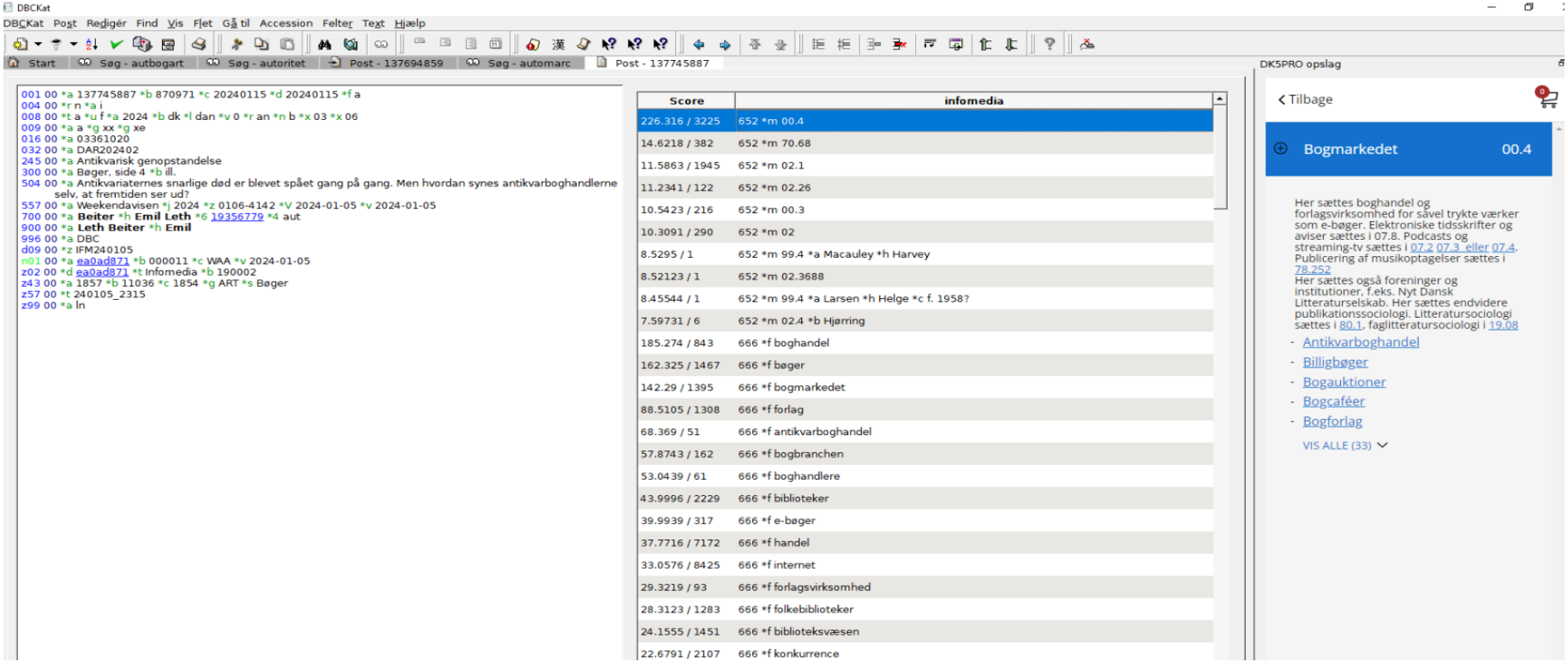
DBC
D1G1TAL

DBCKat



DB
D1G1TAL

DBCKat



DBC
D1G1TAL

Hierarkisk struktur til vores emneord

Udgangspunkt:

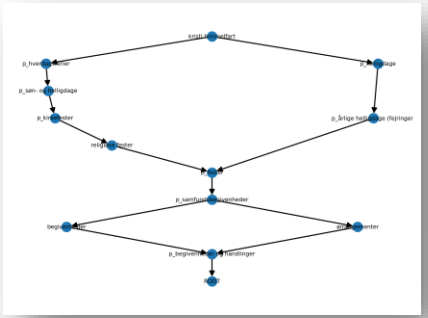
- Til faglitterære fuldtekstartikler inden for Dansk Artikelindeks
- Ca. 40.000 dbc-kontrollerede emneord (i "flad struktur")
- Direkte anvendelse af AI-modeller til fuldtekster for artikler – vi håber vi kan gøre det bedre
- YSO – General Finnish Ontology (National Library of Finland)
- YSO-sprog: finsk, svensk, engelsk

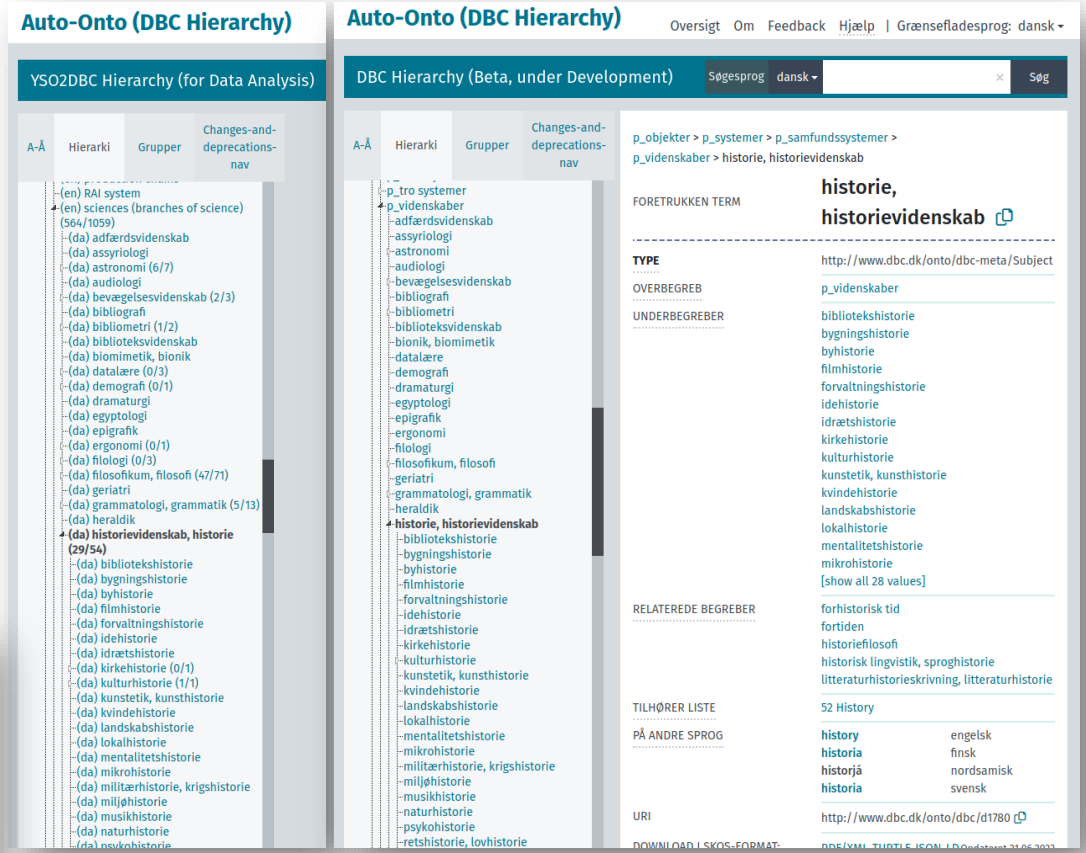
DBC D1G1TAL

https://finto.fi/yso/en/

Udvikling af vores egen ontologi
auto-onto service

- One-to-many mapping
- Tilføjelse af nye overbegreber (f.eks. historiske begivenheder)





DBC D1G1TAL

Mapping af emneord (DBC/danske emneord -> YSO)

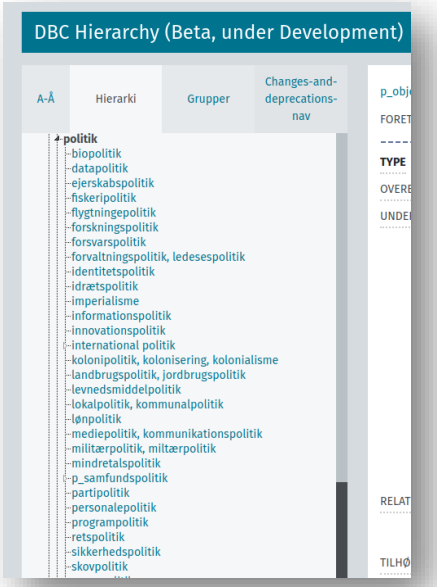
Det er ikke trivielt:

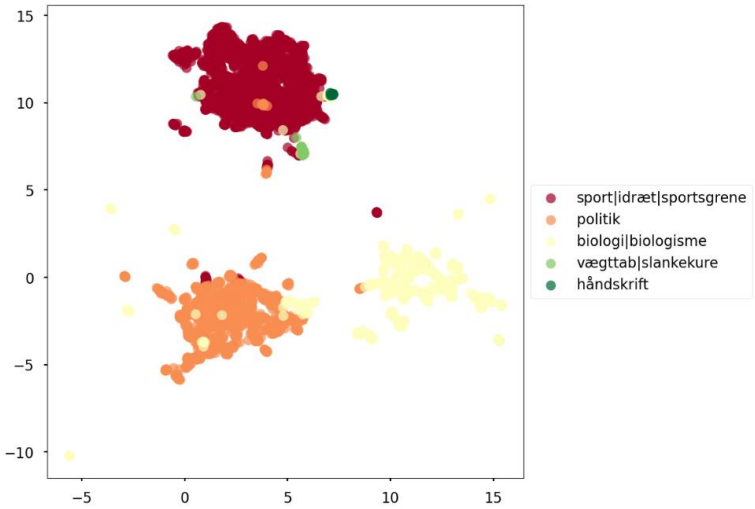
- Biller – programmeringsfejl (bugs)
- Solsejl - markise (marquess)
- Pushere - tryk (pressure)
- Isbjørnejagt => jagt

http://auto-onto-web.mi-staging.svc.cloud.dbc.dk/ (virker kun hos DBC)

Repræsentation af ontologien

Emneord tæt på hinanden i ontologien skal også ligge tæt på hinanden i et vektor space



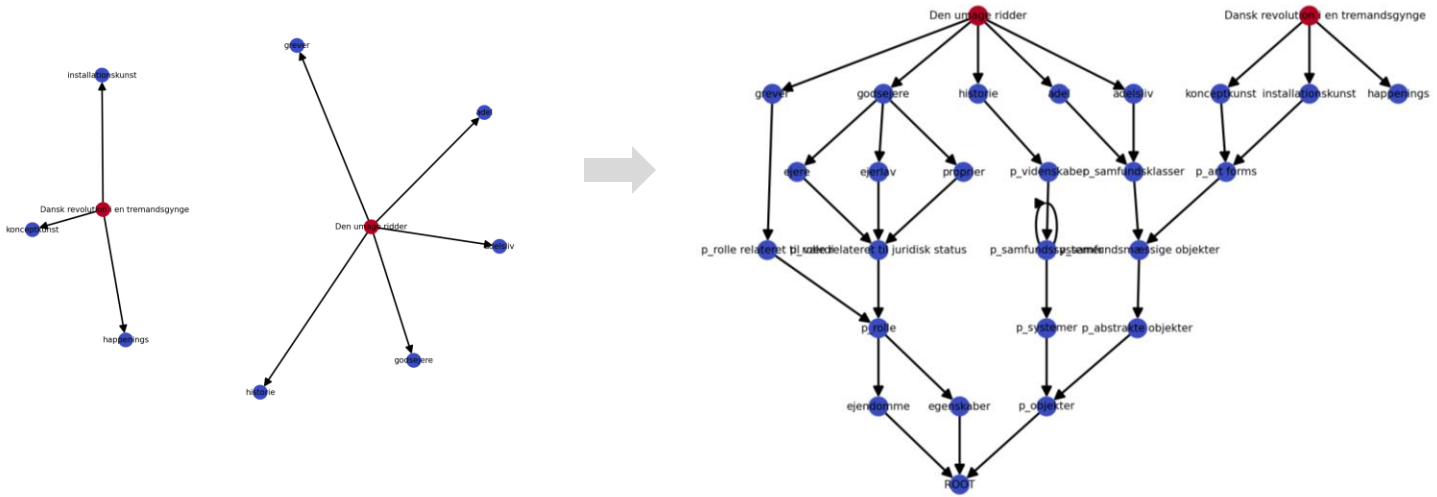


DBC D1G1TAL

Lær en (computer) model til artikelindeksering

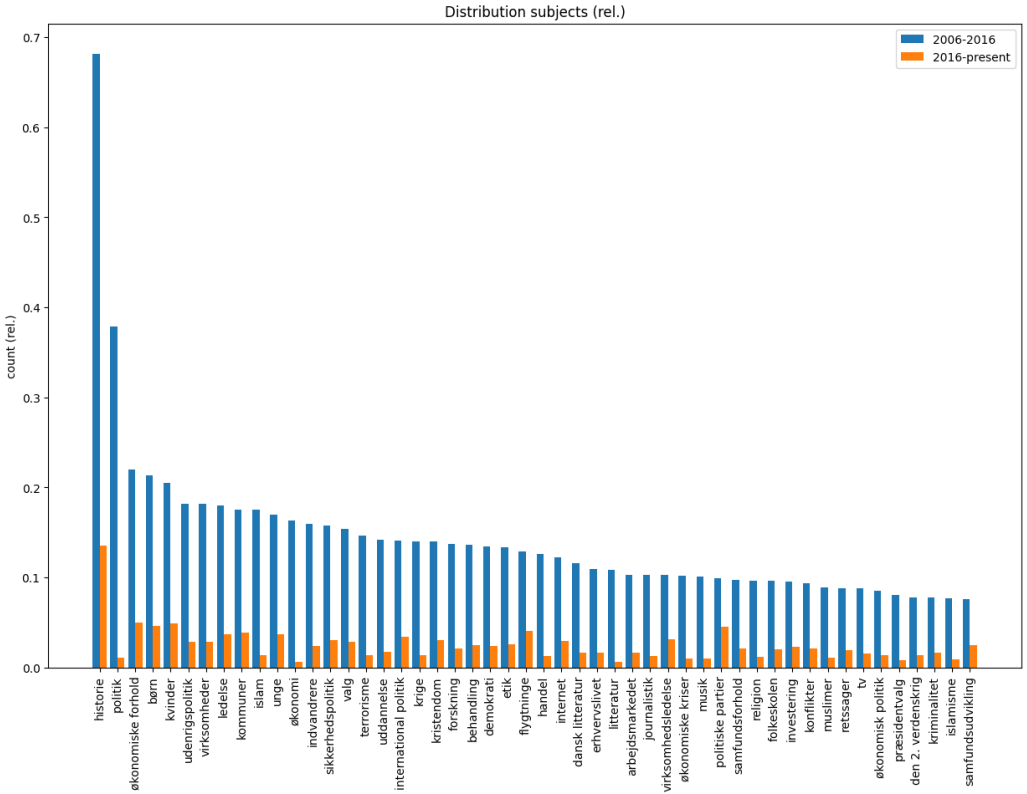
Baseret på eksisterende artikler

Artikler placeres i ontologien og kan sættes i kontekst med hinanden*:



*Note: hierarkiet er i udviklingsfasen og relation fremvist her er ikke direkte meningsgivende mellem de 2 artikler, men skal visualisere muligheden for at extrahere relationer mellem artikler på en overordnet niveau ud af ontologien, f.eks. forbinder en artikel som handler om "atompolitik" med en artikler som handler om "atomvåbenpolitik"

Analyser af emneords-fordeling over alle artikler



DBC
D1G1TAL

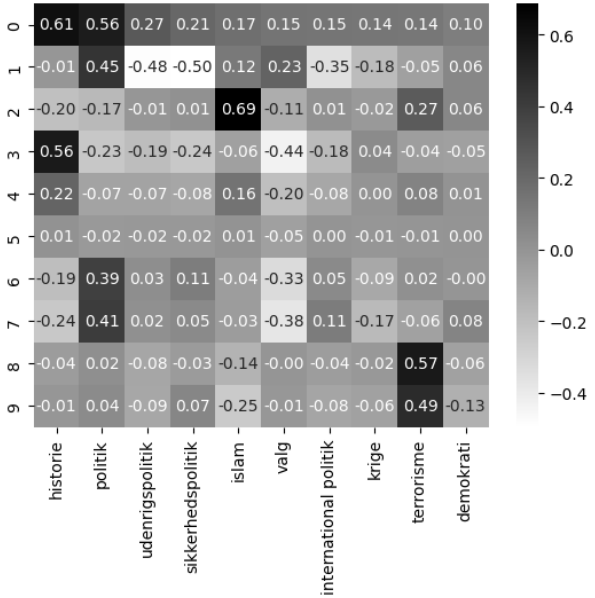
Fra artiklernes indhold til emner - 1

Artikeleksempel:

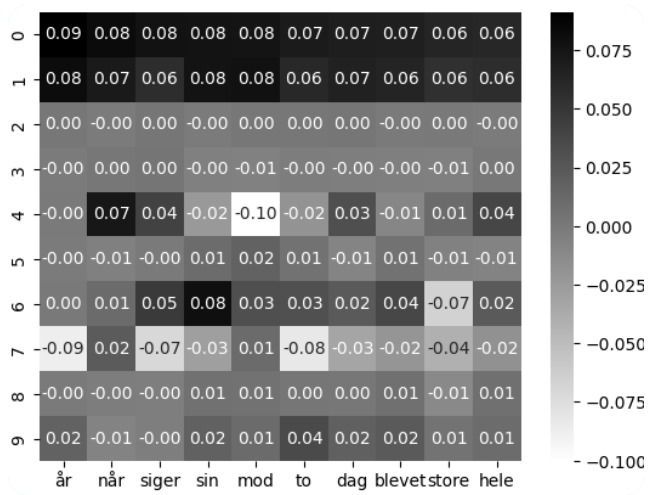
<Heading> Afspænding men ikke sikkerhed
<Body> Europas forsvar er i støbeskeen. Den kolde krig er forbi, og det sikkerhedspolitiske mønster, vi har kendt i fire årtier, eksisterer ikke mere. ….

<Emneord> 'udenrigspolitik','sikkerhedspolitik'

Manuel emneordtildeling:



Automatisk emnefinding:



DBC D1G1TAL

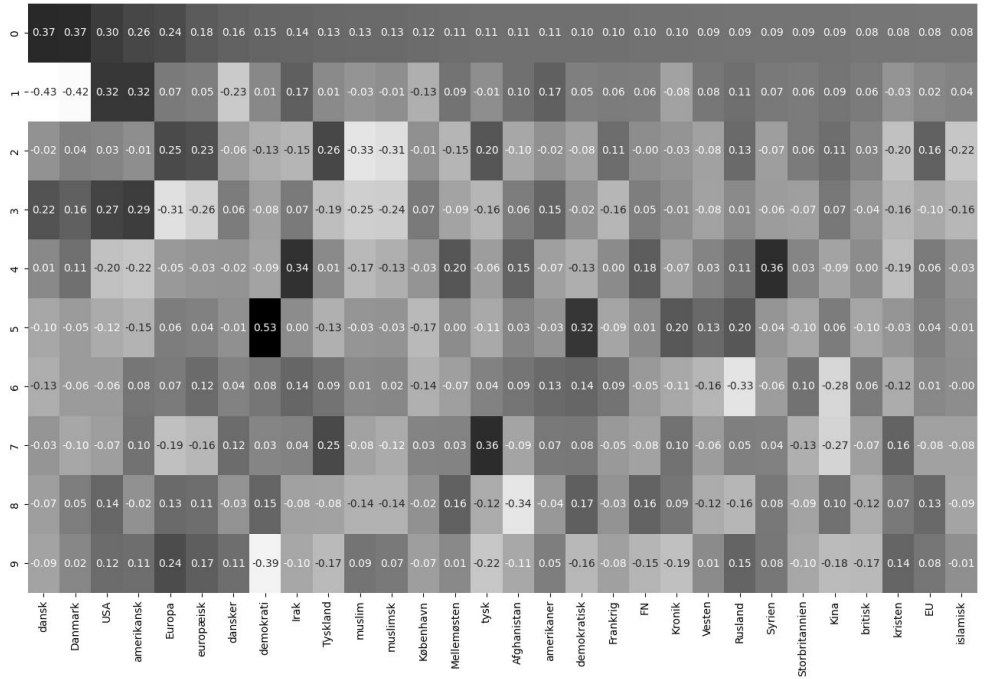
Fra artiklernes indhold til emner – 2 – eks. NER

Artikeleksempel:

<Heading> Afspænding, ikke sikkerhed
<Body> Europas forsvar, Den kolde Krig, sikkerhedspolitiske mønster

<Emneord> 'udenrigspolitik', 'sikkerhedspolitik'

When Sebastian Thrun PERSON started working on English LANGUAGE self-driving cars at Google ORG in 2007 DATE , few people outside of the company took him seriously.
"I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun PERSON ,
now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE .
The interview took place in German NORP because of Thrun's past.
A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand,
to secure their place in the fast-moving world of fully automated transportation.



DBC
D1G1TAL

Opsamling af det tekniske perspektiv

Afgørende for at udvikle Auto-Index:

- Tilgængelige metadata omkring artiklerne som en ontologi
- Datakvalitet af udtagelige informationer

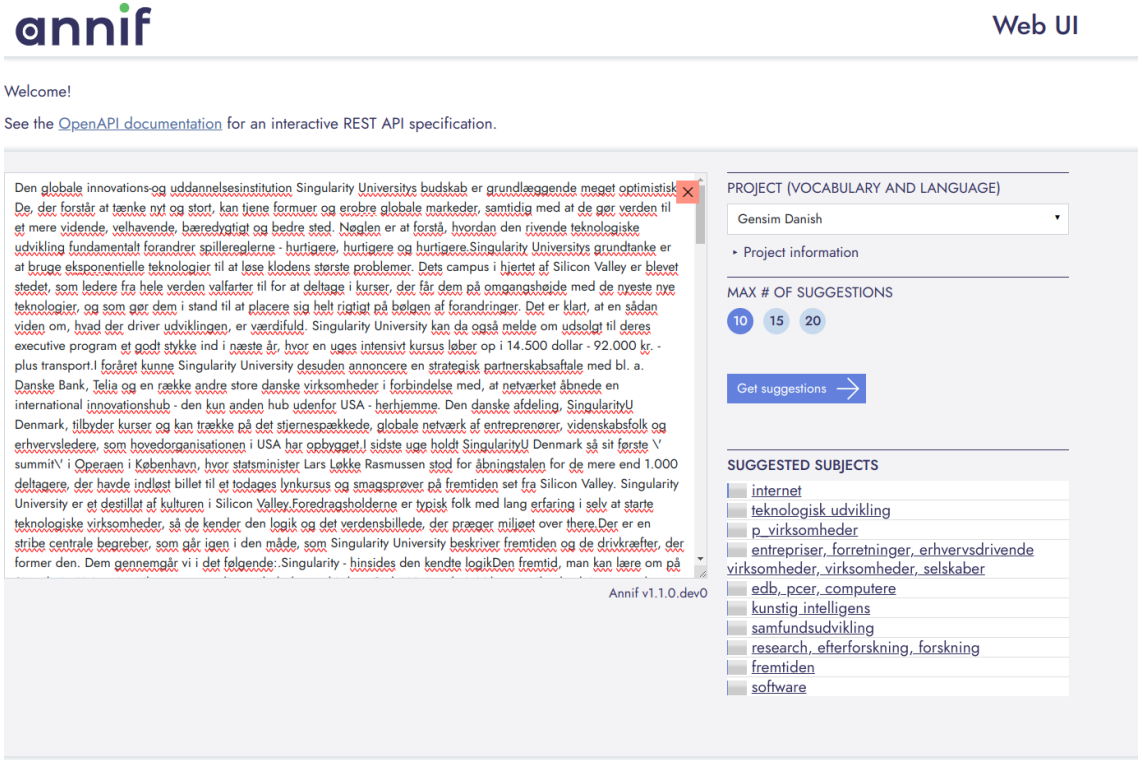
- (Datamængde)

ANNIF

Annif er et framework lavet af de samme folk som startede YSO

- Framework til at bruge sprogmodeller til at udlede emneord fra fuldtekster på forskellige sprog
- Potentielt nyttigt hvis vi vil indlede samarbejde med dem.
- 3-4 modeller indbygget og mulighed for at kombinere
- Vi har tilpasset den eksisterende model til udledning af emner og implementeret den i Annif-frameworket
- Man skal "blot" implementere tre ting:
  - Sæt op (load model og evt andre ting)
  - Træn på et korpus
  - Funktion til at udlede emneord givet en fuldtekst

ANNIF demo (screenshot hvis wifi er dumt)



DBC
D1G1TAL

Future Work:

- Metrik for "godt nok" - eksisterende data, manuelle kontroller
- Change Management i ontologi-ændringer (dynamisk hierarki)
- Implementering i produktionssystemer (arkitektur, overvågning,..)

Perspektiv for Ontologi / Auto-Index

- Automatisk tildeling af emneord til artikler fra kilder uden manuel inspektion
- Bedre søgning - "genberegne" yderligere emneord til bestående artikler
- Navigation / interaktiv søgehjælp ved hjælp af hierarki
- Længerevarende udforskning af synenergieffekt mellem AI-baserede metoder og et dybt domænekendskab (human-expert-in-the-loop)
  - Analyse af hvilke emner vi foreslår der bliver brugt af katalogisatorer og hvilke der bliver forkastet
  - Analyse af hvilke emner der bliver "fundet frem" ved søgninger: de manuelt valgte eller dem der kommer fra modellen?
  - Kan dybt kendskab til domæne hjælpe os til at udvikle bedre modeller?

Tak

https://www.dbc.dk/

Noah Torp-Smith – nots@dbc.dk
Philine Zeinert – phiz@dbc.dk