

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

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

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 ,

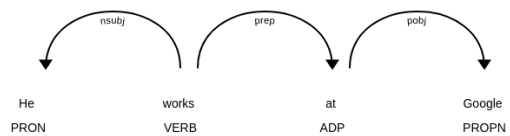
now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE .

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

Brug af metadata i vores AI-services

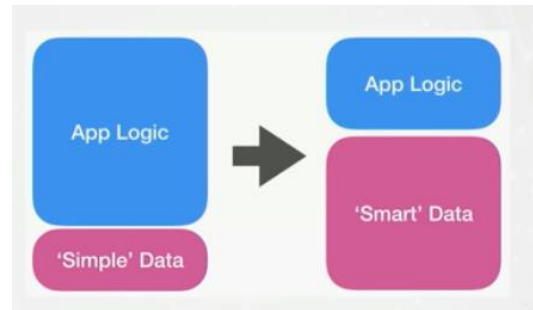
Metadata er over det hele!

- 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



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/materialer/personer etc, så kan vi give bedre søgeresultater.

Appeldata

Brug af stemningsdata i vores AI-Services

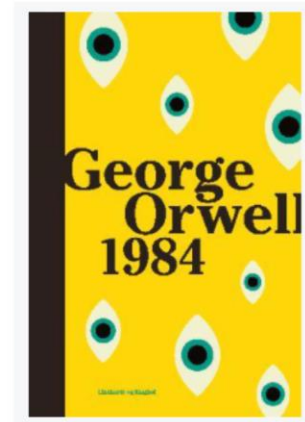
George Orwell - 1984

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
<ul style="list-style-type: none">• upsetting• gloomy• philosophical• thought-provoking	<ul style="list-style-type: none">• Winston Smith• Big Brother	<ul style="list-style-type: none">• plain language	<ul style="list-style-type: none">• 3rd person narrator	<ul style="list-style-type: none">• lingering

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

SLIDERS	SLIDERS	SLIDERS
LET/SVÆR 1-5	VIRKELIG/FANTASI 1-5	KORT/LANG 1-5
Lettilgængeligt sprog: 1 Overskuelig: 1-3 Almindeligt sprog: 2-3 Mange metaforer: 3-5 Handling på flere niveauer: 3-5 Mange detaljer: 3-5 Mange facts: 3-5 Detaljeret sprog: 4-5 Mange indskudte sætninger: 4-5 Krævende sprog: 5	Eksempler: 1: Mira 2: Wimpy Kid 3: Harry Potter (sammenhæng med den virkelige verden) 4: Drageherren 5: Fuld fantasy verden ("High Fantasy")	1 - 1-50 sider 2 - 51-75 sider 3 - 76-100 4 - 101-200 5 - 201+



Type	Emne	Emne	Emne	Emne	Emne
Tag	Dyr	Sport	Venskaber	Mit liv	Ud i fremtiden
Supplerende ord	ikke fantasivæsenner			familien – skolen – hverdag	dystopi – rummet – science fiction
Emoji					

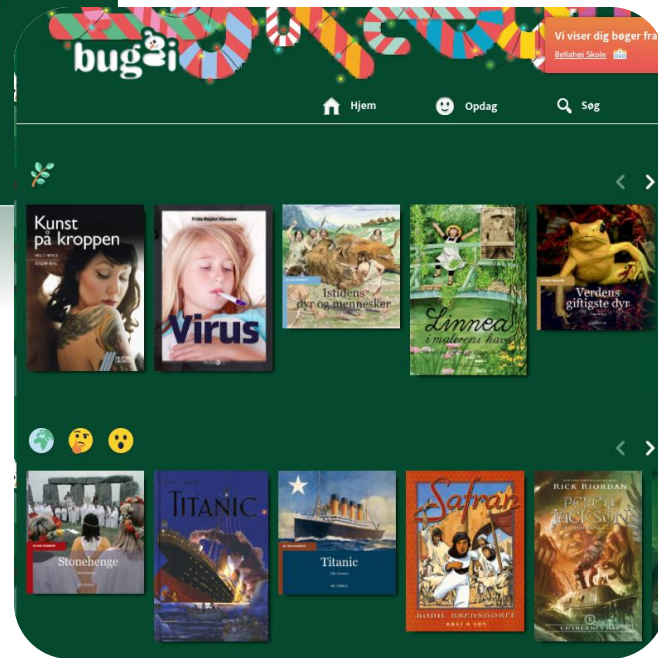
Buggi

Brug af metadata i buggi:

- Søgemaskine
- Suggesterfunktion
- **Anbefalinger**



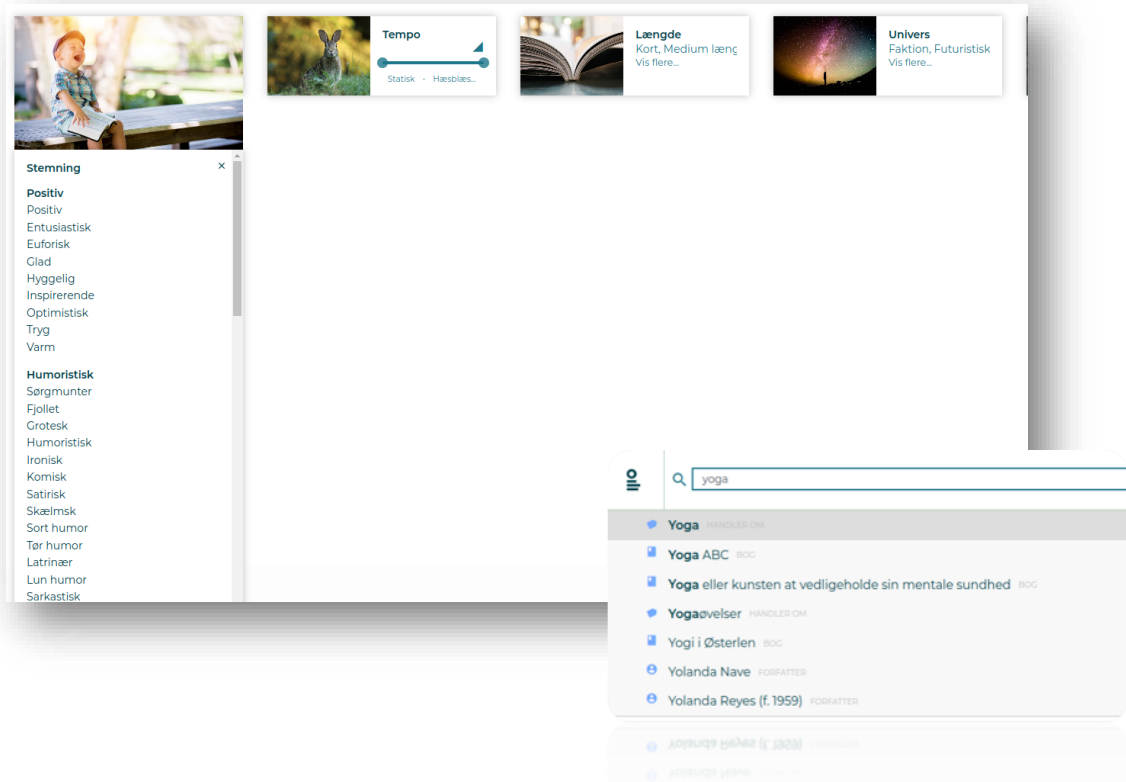
<https://buggi.dk/school>



Appeldata Voksne

Voksne stemningsdata består af:

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



The screenshot displays the Læsekompas app interface. At the top, there are three filter cards: 'Tempo' with a slider set to 'Statisk', 'Længde' with options 'Kort, Medium længde' and 'Vis flere...', and 'Univers' with options 'Fiktion, Futuristisk' and 'Vis flere...'. Below these is a 'Stemning' (Mood) dropdown menu with the following options: Positiv, Entusiastisk, Euforisk, Glad, Hyggelig, Inspirerende, Optimistisk, Tryk, Varm, Humøristisk, Sørgmunter, Fjøllet, Grotesk, Humøristisk, Ironisk, Komisk, Satirisk, Skælmisk, Sort humor, Tør humor, Latrinær, Lun humor, and Sarkastisk. A search bar at the bottom contains the text 'yoga'. Below the search bar, a list of search results is shown, including 'Yoga' (HANDLER OM), 'Yoga ABC' (BOG), 'Yoga eller kunsten at vedligeholde sin mentale sundhed' (BOG), 'Yogaøvelser' (HANDLER OM), 'Yogi i Østerlen' (BOG), 'Yolanda Nave' (FORFATTER), and 'Yolanda Reyes (f. 1959)' (FORFATTER).

Læsekompas

Brug af metadata i læsekompas:

- Søgemaskine
 - Suggester
 - Anbefalinger baseret på tags
 - Anbefalinger baseret på titler
- (Minder om)



FOR TRAVLT TIL ROMANER?

[noveller](#) handling på flere niveauer skæbnsvanger

Prøv en novelle



Allegorisk og entusiastisk med poetisk sprog
Fantastisk og foruroligende



Realistisk eksperimenterende litteratur om familien



Tankevækkende, trist og foruroligende
Historier om omsorgssvigt på Lolland

<https://laesekompas.dk/>



MINDER OM BARE ALICE



Realistisk young adult om identitet



Følsom, varm og charmerende

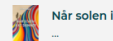


Fortælling om kærlighed i USA's high school



Young adult der udspiller sig i USA

SAMMENLIGNING



Når solen i

vs.



Bare

Læseoplevelse

Fælles læseoplevelse
Kun bogen Når solen ikke skinner

stemning

vemodig

sprog og form

young adult almindeligt sprog meget dialog åben slutning

alvidende fortæller fremadskridende realistisk

handling

følelser kærlighed homoseksuelle lesbiske veninder familien

venskab forelskelse homoseksualitet identitet

tid og sted

Danmark Silkeborg 2010-2019

DBC
DIGITAL

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

```
DBCKat
DBCKat Post Redigér Find Vis Flet Gå til Accession Felter Text Hjælp
Start Søg - autbogart Søg - autoritet Post - 137694859 Søg - automarc Post - 137745887
001 00 *a 137745887 *b 870971 *c 20240115 *d 20240115 *f a
004 00 *r n *a i
008 00 *t a *u f *a 2024 *b dk *l dan *v 0 *r an *n b *x 03 *x 06
009 00 *a *g xx *g xe
016 00 *a 03361020
032 00 *a DAR202402
245 00 *a Antikvarisk genopstandelse
300 00 *a Bøger, side 4 *b ill.
504 00 *a Antikvariaternes snarlige død er blevet spået gang på gang. Men hvordan synes antikvarboghandlerne selv, at fremtiden ser ud?
557 00 *a Weekendavisen *j 2024 *z 0106-4142 *v 2024-01-05 *v 2024-01-05
700 00 *a Beiter *h Emil Leth *6 19356779 *4 aut
900 00 *a Leth Beiter *h Emil
996 00 *a DBC
d09 00 *z IFM240105
n01 00 *a ea0ad871 *b 000011 *c WAA *v 2024-01-05
z02 00 *d ea0ad871 *t Infomedia *b 190002
z43 00 *a 1857 *b 11036 *c 1854 *g ART *s Bøger
z57 00 *t 240105_2315
z99 00 *a ln
```

DBCKat

DBCKat Post Redigér Find Vis Flet Gå til Accession Felter Text Hjælp

Start Søg - autbogart Søg - autoritet Post - 137694859

001 00 *a 137694859 *b 870971 *c 20240105171109 *d 20240105 *fa
004 00 *r n *a i
008 00 *t a *u f *a 2024 *b dk *d y *l dan *n b *r an *x 03 *x 06 *v 0
009 00 *a *g xx *g xe
016 00 *a 03361020
032 00 *a DAR202402
245 00 *a Antikvarisk genopstandelse
300 00 *a Bøger, side 4-5 *b ill.
504 00 *a Antikvariaternes snarlige død er blevet spået gang på gang. Men hvis man spørger antikvarboghandlerne selv, synes de at fremtiden ser strålende ud. Noget tyder dog på, at der kommer til at være færre reoler, mindre støv og færre, men mere kuraterede bøger
557 00 *a Weekendavisen *j 2024 *z 0106-4142 *V 2024-01-05 *v 2024-01-05
652 00 *m 00_4
666 00 *f boghandel
666 00 *f bøger
666 00 *f antikvarboghandel
666 00 *f bogbranchen
666 00 *f boghandlere
666 00 *f bogmarkedet
700 00 *a Beiter *h Emil Leth *6 19356779 *4 aut
900 00 *a Leth Beiter *h Emil
996 00 *a DBC
d08 00 *o ln
d09 00 *z IFM240105
n01 00 *a ea0ad871 *b 000011 *c WAA *v 2024-01-05
s12 00 *t TeamBAM202401
z02 00 *d ea0ad871 *t infomedia *b 190002
z43 00 *a 1857 *b 11036 *c 1854 *g ART *s Bøger
z57 00 *t 240105_1415
z99 00 *a ln

Score	inform
240.498 / 3225	652 *m 00.4
15.3166 / 290	652 *m 02
10.5438 / 216	652 *m 00.3
9.77624 / 382	652 *m 70.68
8.3862 / 1	652 *m 99.4 *a Macauley *h Harvey
7.96733 / 1945	652 *m 02.1
7.78475 / 6	652 *m 02.4 *b Hjørring
7.41863 / 7	652 *m 75.2
7.22879 / 11	652 *m 99.4 *a Thorup *h Emil
7.04431 / 16	652 *m 02.4
203.513 / 843	666 *f boghandel
162.839 / 1467	666 *f bøger
146.623 / 1395	666 *f bogmarkedet
88.513 / 1308	666 *f forlag
81.6171 / 51	666 *f antikvarboghandel
63.2718 / 162	666 *f bogbranchen
53.3517 / 61	666 *f boghandlere
40.4566 / 2229	666 *f biblioteker
37.8668 / 7172	666 *f handel
35.587 / 8425	666 *f internet
34.9916 / 317	666 *f e-bøger
29.312 / 93	666 *f forlagsvirksomhed

DBCKat

DBCKat Post Redigér Find Vis Fjlet Gå til Accession Felter Text Hjælp

Start Sag - autobogart Sag - autoritet Post - 137694859 Sag - automarc Post - 137745887

```
001 00 *a 137745887 *b 870971 *c 20240115 *d 20240115 *f a
004 00 *r n *a i
008 00 *t a *u f *a 2024 *b dk *l dan *v 0 *r an *n b *x 03 *x 06
009 00 *a a *g xx *g xe
016 00 *a 03361020
032 00 *a DAR202402
245 00 *a Antikvarisk genopstandelse
300 00 *a Bøger, side 4 *b ill.
504 00 *a Antikvariaternes snarlige død er blevet spået gang på gang. Men hvordan synes antikvarboghandlerne selv, at fremtiden ser ud?
557 00 *a Weekendavisen *j 2024 *z 0106-4142 *v 2024-01-05 *v 2024-01-05
700 00 *a Beiter *h Emil Leth *g 19356779 *4 aut
900 00 *a Leth Beiter *h Emil
996 00 *a DBC
d09 00 *z IFM240105
n01 00 *a ea0ad871 *b 000011 *c WAA *v 2024-01-05
z02 00 *d ea0ad871 *t infomedia *b 190002
z43 00 *a 1857 *b 11036 *c 1854 *g ART *s Bøger
z57 00 *t z40105_z315
z99 00 *a In
```

Score	infomedia
226.316 / 3225	652 *m 00.4
14.6218 / 382	652 *m 70.68
11.5863 / 1945	652 *m 02.1
11.2341 / 122	652 *m 02.26
10.5423 / 216	652 *m 00.3
10.3091 / 290	652 *m 02
8.5295 / 1	652 *m 99.4 *a Macauley *h Harvey
8.52123 / 1	652 *m 02.3688
8.45544 / 1	652 *m 99.4 *a Larsen *h Helge *c f. 1958?
7.59731 / 6	652 *m 02.4 *b Hjørring
185.274 / 843	666 *f boghandel
162.325 / 1467	666 *f bøger
142.29 / 1395	666 *f bogmarkedet
88.5105 / 1308	666 *f forlag
68.369 / 51	666 *f antikvarboghandel
57.8743 / 162	666 *f bogbranchen
53.0439 / 61	666 *f boghandlere
43.9996 / 2229	666 *f biblioteker
39.9939 / 317	666 *f e-bøger
37.7716 / 7172	666 *f handel
33.0576 / 8425	666 *f internet
29.3219 / 93	666 *f forlagsvirksomhed
28.3123 / 1283	666 *f folkebiblioteker
24.1555 / 1451	666 *f biblioteksvæsen
22.6791 / 2107	666 *f konkurrence

DKSPRO opslag

< Tilbage

Bogmarkedet 00.4

Her sættes boghandel og forlagsvirksomhed for såvel trykte værker som e-bøger. Elektroniske tidsskrifter og aviser sættes i 07.8. Podcasts og streaming-tv sættes i 07.2 07.3 eller 07.4. Publicering af musikoptagelser sættes i 78.252

Her sættes også foreninger og institutioner, f.eks. Nyt Dansk Litteraturselskab. Her sættes endvidere publikationssociologi. Litteratursociologi sættes i 80.1, faglitteratursociologi i 19.08

- Antikvarboghandel
- Billigbøger
- Bogauktioner
- Bogcaféer
- Bogforlag

VIS ALLE (33) ▾

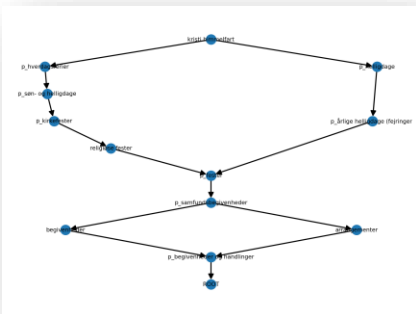
Hierarkisk struktur til vores emneord

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

Udvikling af vores egen ontologi auto-onto service

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



Auto-Onto (DBC Hierarchy)

YSO2DBC Hierarchy (for Data Analysis)

A-Å Hierarki Grupper Changes-and-deprecations-nav

(en) RAI system
• (en) sciences (branches of science) (564/1059)
- (da) adfærdsvidenskab
- (da) assyriologi
- (da) astronomi (6/7)
- (da) audiologi
- (da) bevægelsesvidenskab (2/3)
- (da) bibliografi
- (da) bibliometri (1/2)
- (da) biblioteksvidenskab
- (da) biomimetik, bionik
- (da) datalære (0/3)
- (da) dramaturgi
- (da) egyptologi
- (da) epigrafik
- (da) ergonomi (0/1)
- (da) filologi (0/3)
- (da) filosofikum, filosofi (47/71)
- (da) geriatri
- (da) grammatologi, grammatik (5/13)
- (da) heraldik
• (da) historievidenskab, historie (29/54)
- (da) bibliotekshistorie
- (da) bygningshistorie
- (da) byhistorie
- (da) filmhistorie
- (da) forvaltningshistorie
- (da) idehistorie
- (da) idrætshistorie
- (da) kirkehistorie (0/1)
- (da) kulturhistorie (1/1)
- (da) kunstetik, kunsthistorie
- (da) kvindehistorie
- (da) landskabshistorie
- (da) lokalhistorie
- (da) mentalitetshistorie
- (da) mikrohistorie
- (da) militærhistorie, krigshistorie
- (da) miljøhistorie
- (da) musikhistorie
- (da) naturhistorie
- (da) psykohistorie

Auto-Onto (DBC Hierarchy)

Oversigt Om Feedback Hjælp | Grænsefladesprog: dansk

DBC Hierarchy (Beta, under Development)

Søgning dansk x Søg

A-Å Hierarki Grupper Changes-and-deprecations-nav

p_tro systemer
• p_videnskab
- adfærdsvidenskab
- assyriologi
- astronomi
- audiologi
- bevægelsesvidenskab
- bibliograf
- bibliometri
- biblioteksvidenskab
- bionik, biomimetik
- datalære
- demograf
- dramaturgi
- egyptologi
- epigrafik
- ergonomi
- filologi
- filosofikum, filosofi
- geriatri
- grammatologi, grammatik
- heraldik
• historie, historievidenskab
- bibliotekshistorie
- bygningshistorie
- byhistorie
- filmhistorie
- forvaltningshistorie
- idehistorie
- idrætshistorie
- kirkehistorie
- kulturhistorie
- kunstetik, kunsthistorie
- kvindehistorie
- landskabshistorie
- lokalhistorie
- mentalitetshistorie
- mikrohistorie
- militærhistorie, krigshistorie
- miljøhistorie
- musikhistorie
- naturhistorie
- psykohistorie
- retshistorie, lovhistorie

p_objekter > p_systemer > p_samfundssystemer > p_videnskab > historie, historievidenskab

FORETRUKKEN TERM

historie, historievidenskab

<http://www.dbc.dk/onto/dbc-meta/Subject>

TYPE

OVERBEGREB

p_videnskab

UNDERBEGREB

bibliotekshistorie
bygningshistorie
byhistorie
filmhistorie
forvaltningshistorie
idehistorie
idrætshistorie
kirkehistorie
kulturhistorie
kunstetik, kunsthistorie
kvindehistorie
landskabshistorie
lokalhistorie
mentalitetshistorie
mikrohistorie
[show all 28 values]

RELATEREDE BEGREB

forhistorisk tid
fortiden
historiefilosofi
historisk lingvistik, sproghistorie
litteraturhistorieskrivning, litteraturhistorie

TILHØRER LISTE

52 History

PÅ ANDRE SPROG

history	engelsk
historia	finsk
historjá	nordsamisk
historia	svensk

URI

<http://www.dbc.dk/onto/dbc/d1780>

DOWNLOAD I SYKOS-FORMAT

<http://www.dbc.dk/onto/dbc/d1780>

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

Det er ikke trivielt:

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

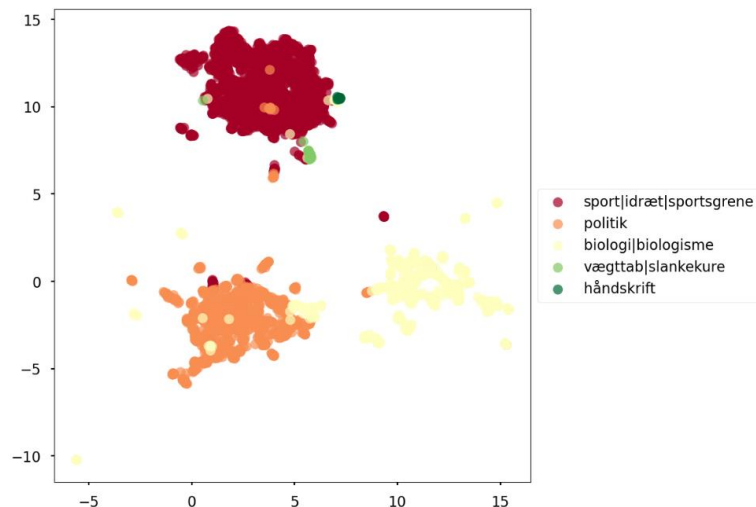
<http://auto-onto-web.mi-staging.svc.cloud.dbc.dk/> (virker kun hos DBC)

Repræsentation af ontologien

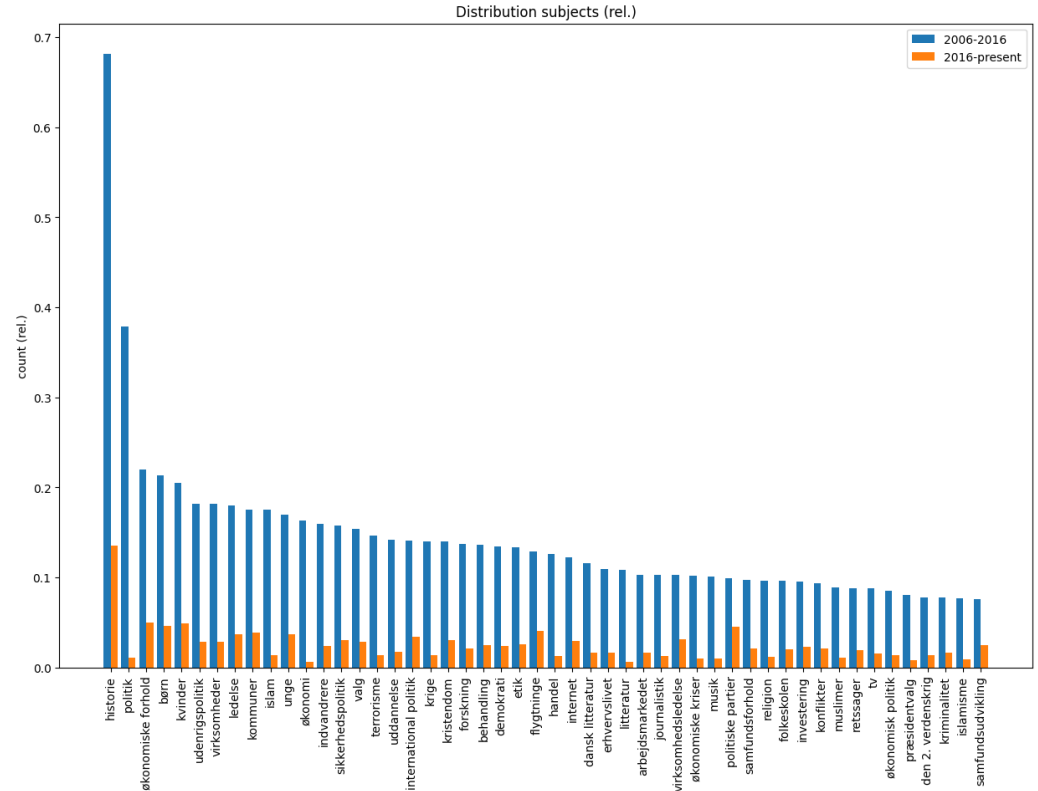
Emneord tæt på hinanden i ontologien skal også ligge tæt på hinanden i et vektor space

DBC Hierarchy (Beta, under Development)

A-Å	Hierarki	Grupper	Changes-and-deprecations-nav	p_obj
	politik			FORET
	-biopolitik			TYPE
	-datapolitik			OVERS
	-ejerskabspolitik			UNDE
	-fiskeripolitik			
	-flygtningepolitik			
	-forskningspolitik			
	-forsvarspolitik			
	-forvaltningspolitik, ledesepolitik			
	-identitetspolitik			
	-idrætspolitik			
	-imperialisme			
	-informationspolitik			
	-innovationspolitik			
	-international politik			
	-kolonipolitik, kolonisering, kolonialisme			
	-landbrugspolitik, jordbrugspolitik			
	-levnedsmiddelpolitik			
	-lokalpolitik, kommunalpolitik			
	-lønpolitik			
	-mediepolitik, kommunikationspolitik			
	-militærpolitik, militærpolitik			
	-mindretalspolitik			
	-p_samfundspolitik			RELAT
	-partipolitik			TILHØ
	-personalepolitik			
	-programpolitik			
	-retspolitik			
	-sikkerhedspolitik			
	-skovpolitik			



Analysér af emneordsfordeling over alle artikler



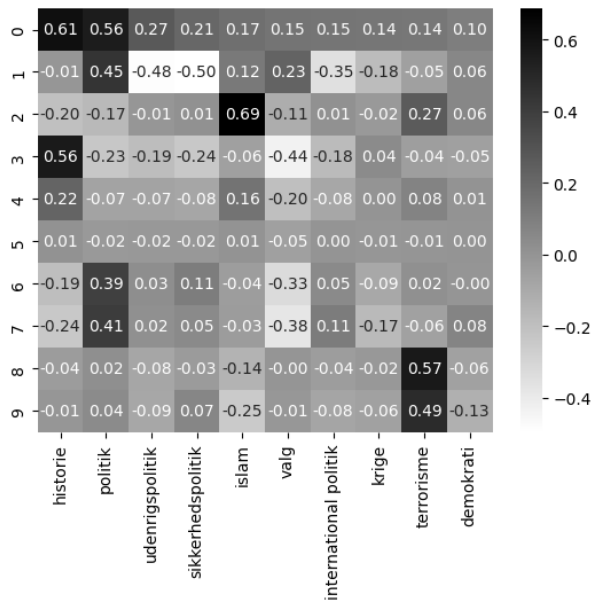
Fra artiklernes indhold til emner - 1

Artikeleksempel:

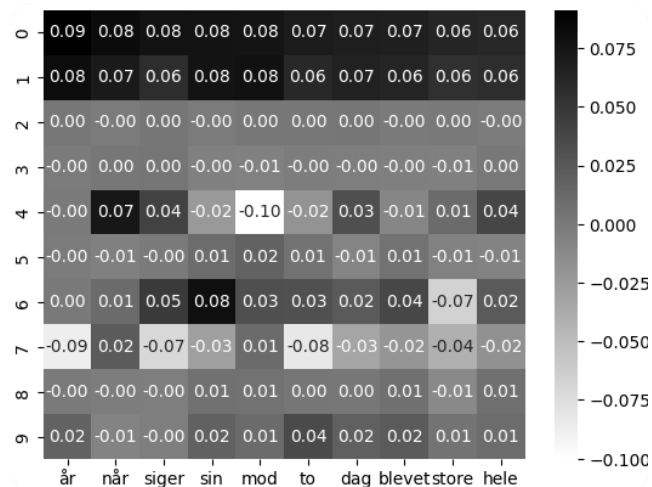
<Heading> Afspænding men ikke sikkerhed
 <Body> Europas forsvar er i støbeskeen. Den kolde krig er forbi, og det sikkerhedspolitiske mønster, vi har kendt i fire årtier, eksisterer ikke mere.

<Emneord> 'udenrigspolitik', 'sikkerhedspolitik'

Manuel emneordtildeling:



Automatisk emnefindning:



Fra artiklernes indhold til emner - 2 - eks. NER

Artikeleksempel:

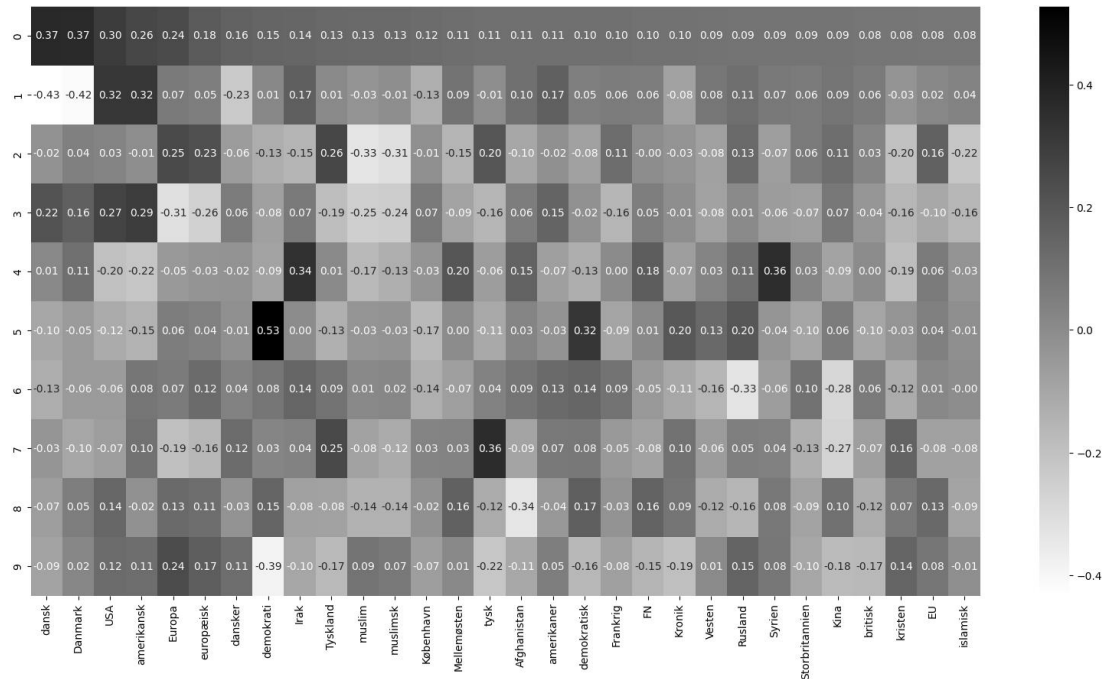
<Heading> Afspænding, ikke sikkerhed
 <Body> Europas forsvar, Den kolde Krig, sikkerhedspolitiske mønster

<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 **Pecode** **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.



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)

annif

Web UI

Welcome!

See the [OpenAPI documentation](#) for an interactive REST API specification.

The screenshot displays the ANNIF demo web interface. On the left, a text analysis window is open, showing a paragraph of Danish text with red underlines indicating detected entities. The text discusses Singularity University's vision and activities. On the right, a sidebar contains a 'PROJECT (VOCABULARY AND LANGUAGE)' section with a dropdown menu set to 'Gensim Danish' and a 'Project information' link. Below this is a 'MAX # OF SUGGESTIONS' section with radio buttons for 10, 15, and 20, and a 'Get suggestions' button. At the bottom of the sidebar is a 'SUGGESTED SUBJECTS' list with items like 'internet', 'teknologisk udvikling', and 'p_virksomheder'. The bottom right corner of the interface shows the version 'Annif v1.1.0.dev0'.

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 – notes@dbc.dk

Philine Zeinert – phiz@dbc.dk