DBC
D1G1TAL

AI til bedre udnyttelse af emneord og som alternativ til manuelt tildelte emneord

Christian Boesgaard, udviklingschef

Introduktion

Jeg har været på DBC siden 2004 og arbejdet med AI siden 2011, bl.a.:

- med til at lave DBCs AI-team
- understøttelse af katalogisering (emneord og DK5)
- recommendersystemer
- (søgning)

Men de sidste par år primært strategi og design af systemer
(så jeg kan være lidt rusten på detaljer og seneste forskning).

Jeg beklager blandingen af dansk og engelsk, men min viden kommer fra engelske kilder, derfor de mange engelske termer.

Kan vi bruge AI til at tildele og udnytte emneord?

- Introduktion til AI, machine learning, natural language processing (NLP) og sprogmodeller
- Muligheder med sprogmodeller og emneord

Send gerne ideer og andet input til mig: cbo@dbc.dk

DBC
D1G1TAL

Introduktion til AI, machine learning, natural language processing (NLP) og sprogmodeller

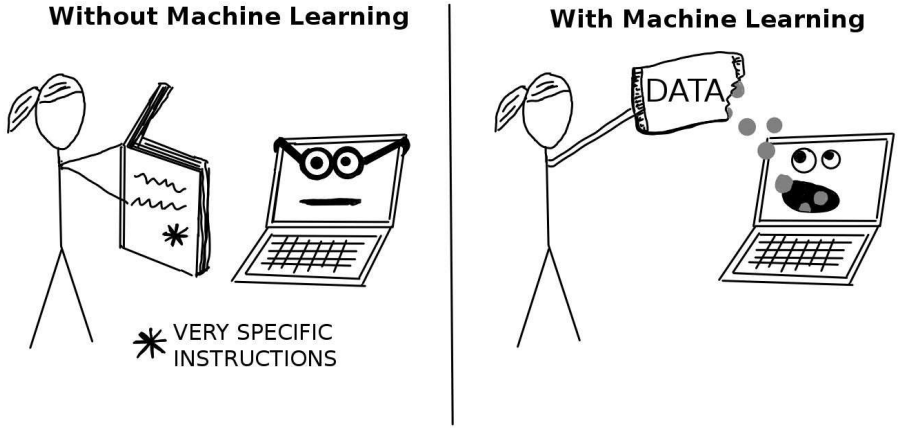
Introduktion til AI – machine learning

Klassisk programmering

- vi opstiller nogle regler og implementerer dem i et program
- fx: udfra en fodboldspillers data - hvor mange mål vil han score i næste sæson?

Machine learning (supervised):

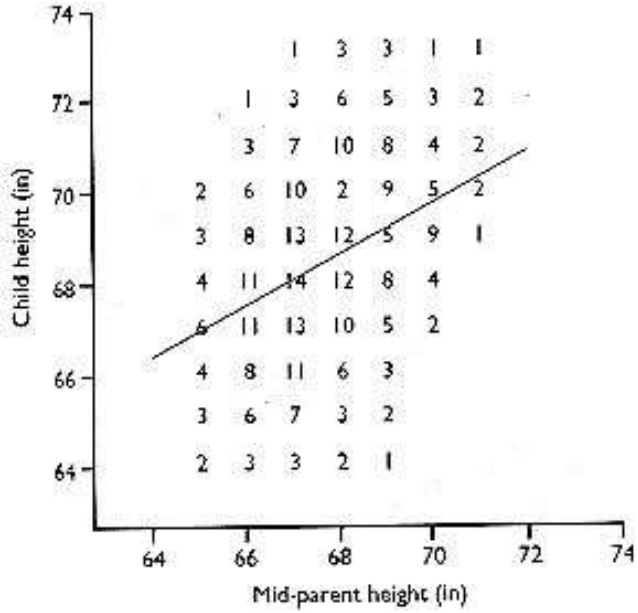
- vi giver nogle eksempler på input og resultat til et system, som giver os et "program"
- "maskinen lærer fra data"
- fx: fodboldspilleres data og mål i den efterfølgende sæson



Introduktion til AI – machine learning

- **linær regression - en klassiker**
- hvor højt bliver et barn hvis vi kender forældrenes højde?
- Galton F. “Regression towards mediocrity in hereditary stature”. *Journal of the Anthropological Institute1886;*15:246–63.

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$



Machine learning elsker meget data og store computere

Machine learning er et fantastisk værktøj til nogle problem og håbløst til andre (“hvad er summen af to tal?”).

- lav et program der kan kende forskel på billeder af bananer og æbler
- ekstremt vanskeligt at løse med klassisk programmering
- kan klares på nogle timer af en studerende på “machine learning 101”

Natural Language Processing (NLP) - language models/sprogmodeller

To tilgange til at modellere naturligt sprog (generelt):

- formal language theory (lingvistik)
- probability theory - herunder sprogmodeller (ML)

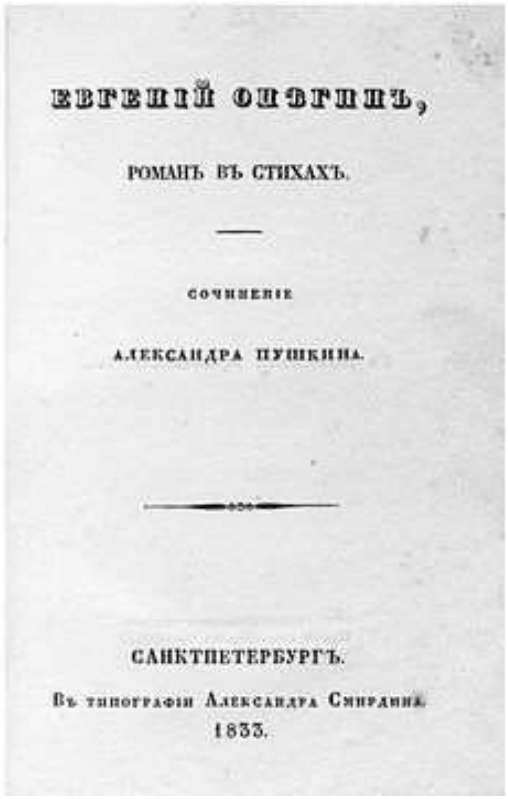
Den første tilgang er elegant, den sidste "virker".

- language models:
  - *"probability distribution on word sequence"*
- neural language models:
  - deep learning based language models
  - deep learning = kæmpestore neurale netværk

God oversigt: Language Models: Past, Present, and Future, Hang Li, Communications of the ACM, July 2022, Vol. 65 No. 7, Pages 56-63 10.1145/3490443

Simpel sprogmodel - Markov chains

- Alexander Markov
- 1913: sandsynlighedsmodel baseret på Alexander Pushkins *Eugine Onegin*
- et ord afhænger af ordene på de sidste n positioner
- model der kan lave tekst der ligner noget et menneske har skrevet


ЕВГЕНІЙ ОНѢГИНЪ,

РОМАНЪ ВЪ СТИХАХЪ.

СОЧИНЕНІЕ

АЛЕКСАНДРА ПУШКИНА.

САНКТПЕТЕРБУРГЪ.

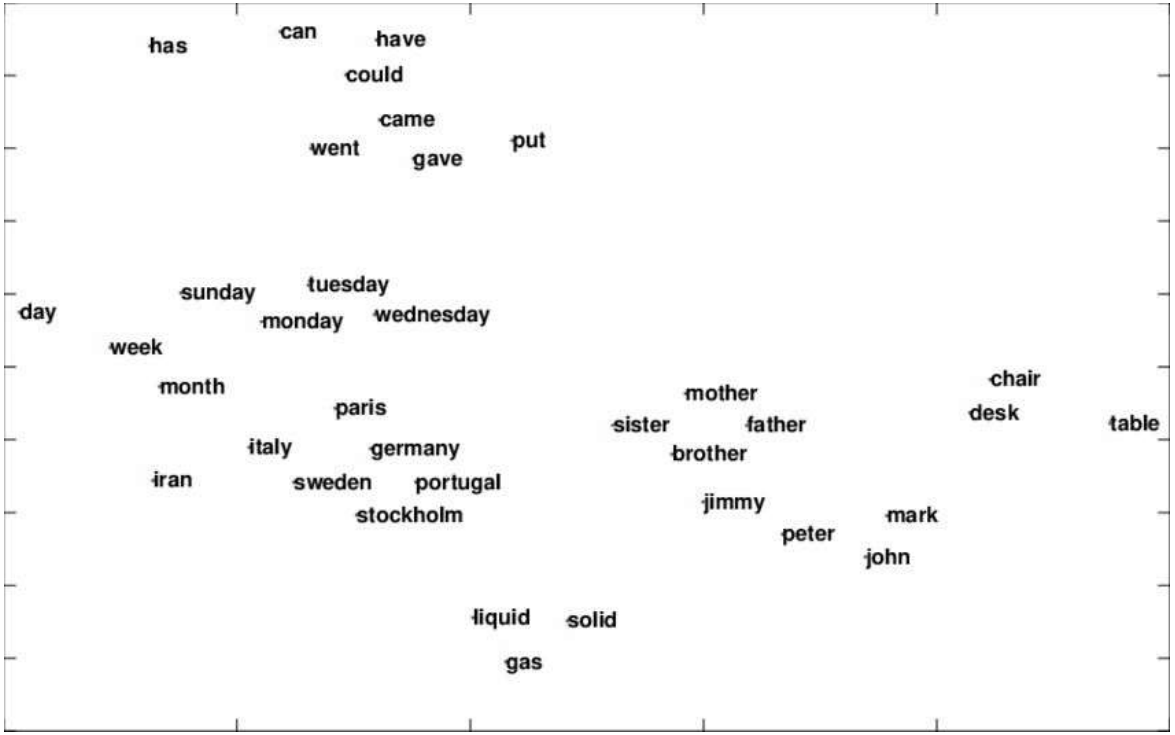
Въ типографіи Александра Смирдина.

1833.

Sprogmodel - eksempel på word embeddings

En sprogmodel der har den egenskab at ord der semantisk ligner hinanden, ligger tæt på hinanden

- figuren viser en sprogmodel, hvor et ord har to dimensioner: “x,y”
- praksis: ord repræsenteres med 2 tal mellem 0 og 1
- 300+ dimensioner er mere realistisk



DBC
D1G1TAL

Hvad kan sprogmodeller bruges til?

- machine translation - oversættelse mellem sprog, state of the art
- reading comprehension - besvare spørgsmål om en tekst
- text generation - fx brugt i sportsnyheder og medicinske svar
- dialogue - chatbot
- search engine - forudsig relevans fra sammenligning mellem søgning og dokument
- **entity extraction** - finde entiteter i tekst, fx geografiske steder
- **similarity** - ligner ord/sætninger/dokumenter hinanden?
- **text summarization** - abstract

Neural language models - eksempler

Hvad mener vi med meget data og store computere?

BERT - Google

- 340 mio. parametre
- trænet på 3.4 mia ord
- Wikipedia har 3 mia. ord

GPT-3 - OpenAI

- 175 mia. parametre
- trænet på 500 mia. ord på over 100 sprog
- estimeret til at have kostet $10-20 mio. i computertid at træne

Vi kan heldigvis bruge eksisterende modeller ved at modificere dem.

GPT-3 eksempel 1

GPT-3 er et eksempel på en model der kan bruges til at lave tekst ud fra et “prompt”, et simpelt eksempel er at fortsætte en sætning:

Input: in the future research librarians will use ai

Output: [...] to help them fulfil their duties. This will include using ai to help cataloguing, to identify trends in research, and to help with reference queries.

GPT-3 eksempel 2

Input: Libraries help students and researchers

Output: [...] by providing them with a place to access information and resources. Libraries also offer a variety of services, such as reference assistance, research help, and educational programming.

DALL·E 2 eksempel 1

Modeller kan være multi-modale, fx arbejde med både tekst og billeder.

DALL-E 2 kan lave billeder ud fra tekst:

Input: research librarians at conference

DALL·E 2 eksempel 2

Input: research librarians at conference as a cartoon

(modeller afspejler bias fra datagrundlag, fx er det næppe tilfældigt at der kun er kvinder på billedet).

DBC
D1G1TAL

Muligheder med sprogmodeller og emneord

Muligheder med sprogmodeller og emneord

Følgende skal ses som et ide-katalog til inspiration

- giver ideerne overhovedet mening?
- kan de være afsæt for bedre ideer?

Og det hele vil kræve:

- eksperimenter og afprøvning - virker det?
- tuning med input fra afprøvning.
- overvejelser om kontekst for brug

Stil gerne spørgsmål til denne del!

Supplement til eksisterende løsninger

Der findes allerede en række løsninger, fx

- Google Scholar
- Dimensions(.ai)
- værktøjer i diverse baser

Vi skal ikke lave noget der allerede findes og måske gør det bedre, men godt med inspiration og hvad kan vi så gøre med vores samlinger og data?

Fem ideer

Søgning og navigation

- Forslag til relaterede emneord
- Søgning direkte i rum med emneord
- Bedre udnyttelse af emneord vha. ontologi

Nye emneord - forudsætning: adgang til fuldtekst

- Emneord fra fuldtekster - ukontrollerede
- Emneord fra fuldtekster - kontrollerede

Kontekst: “det der er i bibliotek.dk”

Legoklods: similarity i sprogmodel

Vi kan sammenligne forskellige emneord – hvor meget ligner de hinanden?

Input: **machine learning og emneord vi ønsker at sammenligne med.**

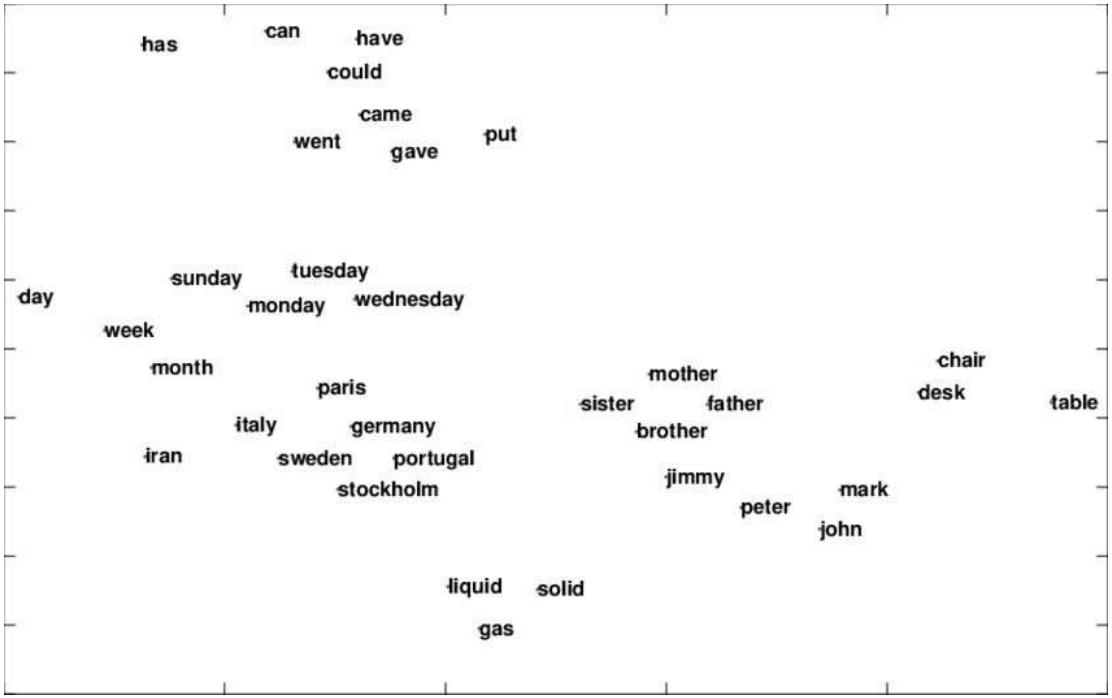
Output:
natural language processing 0.560
computer science 0.552
neural network 0.356
mathematics 0.280
medicine 0.162

Vi ville også istedet kunne bede om emneord (fra mængde) der ligner “machine learning”

Bedre udnyttelse af emneord med similarity

Med udgangspunkt i en egnet sprogmodel kan vi:

- lægge heterogene emneord ind et passende rum
- optimalt en sprogmodel der understøtter flere sprog i samme rum



DBC
D1G1TAL

Forslag til relaterede emneord

Udfra emneord der ønskes søgt efter kan vi foreslå relaterede emneord der kan søges på, dvs. søgning kan bredes ud (OR-søgning) til at dække forskellige emneord der beskriver det samme.

Det er stadig en person der vælger den endelige søgning og den forventede høje precision med emneord opretholdes.

Eksempel: “language models” kan udvides med delmængde af forslag: “neural language model”, “embedding model”, “BERT”, “GPT-3”

Søgning →forslag til udvidelse af søgning →manuelt valg → udvidet søgning

Søgning direkte i rum med emneord

Udfra en søgning kan vi finde relaterede emneord i rummet og automatisk bruge disse til at lave en søgning.

Dette forventes at give en bredere søgning (med tab af precision).

Søgning →automatisk udvidelse af søgning →udvidet søgning

Bedre udnyttelse af emneord vha. ontologi

Med udgangspunkt i en egnet sprogmodel kan vi:

- oversætte emneord til en eksisterende ontologi
- fx Wikipedia eller "YSO - General Finnish ontology"

Dette muliggør navigation i samling via ontologi, dvs. fra en placering i ontologien kan der laves søgninger på eksisterende emneord – enten manuelt med valg af ønskede emneord eller automatisk.

Navigation i ontologi →forslag til søgning → manuelt valg → søgning

Navigation i ontologi →automatisk konstrueret søgning

Emneord fra fuldtekster - ukontrollerede

Vi kan udtrække emneord fra fuldtekster (det kunne man også før moderne NLP, men det kan nu gøres lidt bedre).

Disse kan bruges i sædvanlige søge- og visningssituationer.

Fordelen ved at udtrække emneord der er i teksterne er en forventet høj precision på søgninger. Kan ses som selektiv fuldtekst-indeksering.

Ulempen er et meget højt forventet antal emneord – forventede udfordringer kan dog afhjælpes med tidligere præsenterede metoder til søgning.

Fuldtekst →ukontrollerede emneord der findes i fuldtekst

Emneord fra fuldtekster - kontrollerede

Vi kan tildele emneord (fx) fra en ontologi ved at placere fuldteksten et eller flere steder i ontologien. Eller udtrække ukontrollerede emneord og “oversætte” dem til ontologi.

Emneordene kan bruges i sædvanlige søge og visningssituationer.

Fordelen ved den tilgang er et kontrolleret sæt af emneord der kan gøre søgning nemmere - ligesom ontologien kan bruges direkte til navigation af emneord.

Ulempen er et forventet tab af precision, pga. emneord der ikke rammer helt rigtigt.

Fuldtekst →kontrollerede emneord fra ontologi

DBC
D1G1TAL

The end

Send gerne ideer og andet input til mig: cbo@dbc.dk