Matematik, maskiner og metadata

af CHRISTIAN BOESGAARD
DATALOG
IT Development
/ DBC

Konkrete projekter med machine learning, hvor computersystemer lærer fra data, har allerede udmøntet sig i systemer, der kan assistere den manuelle indeksering. Perspektiverne er, at man kan give metadata til materiale, der ellers ikke ville have fået det, og at mennesket og maskinen formentlig engang i fremtiden kommer til at arbejde sammen på en ny måde.

I disse år er der stor opmærksomhed på, hvad man kan få ud af at analysere data. IBM har med Watson-projektet demonstreret, at maskiner kan vinde over mennesker i Jeopardy, Google har konstrueret en selvkørende bil, og Barack Obamas seneste præsidentkampagne var i høj grad baseret på analyser og forudsigelser lavet med udgangspunkt i statistik og store datamængder.

De tre umiddelbart forskelligartede eksempler har det til fælles, at de er baseret på at finde mønstre i data og bruge dem til at skabe værdi.

På DBC har vi spurgt os selv, om vi ikke også kan bruge nogle af de samme metoder. DBC har metadata på millioner af materialer, data om hvilke materialer der udlånes sammen, data om hvordan brugere søger i bibliotek.dk, og adgang til fuldteksten for en del af materialerne.

I en gruppe, på tværs af afdelingerne Data og IT Development, har vi i et års tid arbejdet med at inddrage statistik og machine learning til at hjælpe med at skabe mere og bedre metadata og tilbyde biblioteksbrugerne bedre søgemuligheder.

Analyse af eksisterende metadata

En af DBC's hovedopgaver er at udarbejde den danske nationalbibliografi og en brugerrettet bibliotekskatalogisering. Formålet er at gøre det muligt for brugere af folke-, uddannelses- og forskningsbibliotekerne at søge i materialesamlingerne, herunder på bibliotek.dk.

Der bliver katalogiseret 45.000-50.000 materialer om året til Nationalbibliografien, hvoraf ca. halvdelen også bibliotekskatalogiseres, og dermed beskrives mere detaljeret med metadata, eksempelvis emneord og klassemærke (DK5-klassemærke).

Kvaliteten af metadata er afgørende for søgemulighederne, og tildelingen af emneord og klassemærke – såkaldt indeksering – er en krævende manuel opgave, der foretages af specialister. Derudover foregår der en løbende manuel evaluering og kvalitetssikring, der skal sikre, at metadata bruges konsistent og præcist.

Med det formål at gøre det nemmere at tildele og evaluere metadata har vi analyseret brugen af emneord og klassemærke for det samlede sæt af metadata for faglitteratur. Vi har gjort sammentællinger og aggregeringer tilgængelige i et system, hvor man direkte kan se på samforekomster af emneord og klassemærker for alle poster.

Man kan eksempelvis slå emneordet 'knive' op og se, at det oftest forekommer sammen med emneordet 'fremstilling', men også ofte sammen med 'våben' eller 'våbenlovgivning'. Man kan også se, at 'knive' oftest er emneord på poster, der har klassemærkerne 62.5 (Værktøj. Værktøjsmaskiner) eller 62.623 (Blankvåben).

"Hvis man antager, at der er en sammenhæng mellem teksten i et materiale og metadata for materialet, kan man så ikke få computere til – på baggrund af teksten – at foreslå metadata?"

I første omgang er det målet, at systemet skal indgå som et simpelt hjælpeværktøj i forbindelse med manuel indeksering, f.eks. til at undersøge brug af et emneord eller til at foreslå andre emneord eller klassemærker. Det nuværende system er en webgrænseflade, men det er målet, at systemet skal integreres i det eksisterende katalogiseringsværktøj. På længere sigt forestiller vi os, at systemet skal kunne hjælpe med at give en mere omfattende og præcis evaluering af katalogiseringspraksis. Det vil ske ved en gradvis udvidelse og tilpasning af systemet i takt med, at det benyttes.

Fra fuldtekst til automatisk metadata

En del materialer er i dag tilgængelige som fuldtekst, og da skabelsen af metadata er en manuel og ressourcekrævende opgave, er det et oplagt sted at forsøge at få hjælp fra matematik og maskiner: Hvis man antager, at der er en sammenhæng mellem teksten i et materiale og metadata for materialet, kan man så ikke få computere til – på baggrund af teksten – at foreslå metadata? Det kunne bruges som hjælp ved indeksering men også til automatisk at skabe detaljeret metadata på noget af den halvdel af Nationalbibliografien, der ikke bliver bibliotekskatalogiseret.

Vi har derfor udviklet et system baseret på machine learning, der tager udgangspunkt i simple modeller

for nemt at kunne forstå og tilpasse systemet. Vi har arbejdet med fagtekster på artikelområdet, og systemets opgave er at foreslå emneord og klassemærke på baggrund af fuldtekster. Datagrundlaget er et antal artikler i fuldtekst, der allerede er tildelt metadata.

Som eksempel kan vi tage artiklen 'EU-regler gennemhuller lærlingeregler' af Rikke Brøndum i Berlingske Tidende den 25. november 2013, hvor første afsnit lyder:

"Politikere fra både regeringspartier og kommuner har gjort det til en mærkesag at stille krav om lærlinge på offentlige byggerier lige fra skoler til kommende milliardprojekter som Femern Bælt-forbindelsen. Men stik mod hensigten om at uddanne danske lærlinge kan klausulerne i stedet bruges til at uddanne andre landes arbejdskraft. Udenlandske virksomheder kan nemlig besætte pladserne med deres egne elever, fordi EU forbyder krav om nationalitet i offentlige udbud. I nogle tilfælde vil virksomhederne tilmed kunne oprette lærlingepladserne i deres hjemland."

Til denne artikel foreslår vores system emneordene: EU-udbud, anlægsbranchen, byggebranchen, klausuler, licitation, lærlinge, offentlige udbud, praktikpladser og udbud. Og klassemærkerne 33.11 (Arbejdsforhold) eller 33.115 (Arbejdskonflikter).

Ideen bag systemet bygger på den simple antagelse, at tekster, der ligner hinanden, har samme metadata. For at foreslå metadata til en tekst, finder systemet derfor tekster i datagrundlaget, der ligner denne, og foreslår så de samme metadata, som disse har.

Da systemet tager udgangspunkt i de metadata, der er tildelt artiklerne i datagrundlaget, er det vigtigt, at datagrundlaget løbende opdateres (manuelt), f.eks. hvis nye emneord bliver taget i brug.

I systemet repræsenteres teksterne af en delmængde af de ord, der indgår i teksterne, og teksterne sammenlignes baseret på de ord, de har til fælles. Ordene er vægtet ud fra en statistisk analyse af alle ord i alle de tekster, der indgår i datagrundlaget, og vægtene indgår i sammenligningen.

Mere konkret er teksterne repræsenteret af såkaldte vektorer, der er lange rækker af tal, hvor placeringen i rækken angiver, hvilket ord, der repræsenteres, og tallet er en vægtning af ordets betydning. Hvis ordene "EU forbyder krav om nationalitet" indgik i en tekst, kunne plads nr. 323.836 i den tilsvarende vektor repræsentere 'EU' og have værdien 0,139320, og plads 853.773 repræsentere 'nationalitet' og have værdien 0,078814. Når systemet skal foreslå metadata til en tekst, starter det med at konstruere en vektor, der svarer til teksten, og den sammenlignes så med de vektorer, der svarer til teksterne i datagrundlaget. Metadata foreslås så ud fra de vektorer, der ligner mest.

Vores foreløbige resultater viser, at resultaterne er brugbare, men at kvaliteten ikke overraskende er lavere end for manuelt producerede metadata. Det er ikke muligt at lave en præcis maskinel evaluering, da tildelingen af metadata i sidste ende er en subjektiv proces, hvor der kan være flere rigtige muligheder. En mere udførlig vurdering vil derfor først komme i forbindelse med brug af systemet. Vi starter med at tage systemet i brug som et supplerende værktøj for de medarbejdere, der katalogiserer.

Systemet skal integreres i katalogiseringsværktøjet på sigt, men det er allerede tilgængeligt som et selvstændigt system, således at man kan se forslag til emneord og klassemærke på artikler, der skal indekseres.

Fremadrettet vil vi forbedre systemet ud fra de erfaringer, vi får fra praktisk brug, men vi vil også afprøve mere avancerede machine learning-tilgange med en forventning om at forbedre resultaterne.

Som tidligere nævnt vil det være oplagt at tage systemet i brug til den del af nationalbibliografien, som ikke bibliotekskatalogiseres, og hvor der findes fuldtekster. Men en mere avanceret løsning til automatisk tildeling af metadata kan åbne nye muligheder, sådan at det bliver muligt at give metadata til praktisk talt ubegrænsede antal materialer, f.eks. alle danske online-medier eller websider.

Det kan nævnes, at hvis dette kapitel behandles af vores system, så bliver der bl.a. foreslået emneordene: biblioteker, folkebiblioteker, emneord, informationssøgning, klassifikation og søgemaskiner, og

Machine learning – når systemer lærer fra data

Machine learning er et område inden for datalogi og kunstig intelligens og har også stærke bånd til statistik.

Der findes forskellige former for machine learning. Den form, vi har benyttet til at foreslå metadata for tekster, kaldes supervised learning.

Forenklet sagt, så er ideen, at man 'fodrer' systemet med data i form af en række 'opgaver' og deres 'løsning'. Herudfra generaliserer systemet sig så frem til en metode til siden at kunne løse nye opgaver af lignende slags.

Hvis man f.eks. giver systemet en mængde tekster (opgaverne) og emneord (løsningerne), er målet, at systemet på et tidspunkt selv kan regne ud, hvilke emneord, der vil være gode løsninger til nye opgaver – altså til tekster, som systemet ikke tidligere har set.

Det kan naturligvis kun lade sig gøre at generalisere fra data, hvis der findes en underliggende sammenhæng mellem input (opgaver) og output (løsninger). En af udfordringerne ved at arbejde med machine learning er at forstå data og de komplekse sammenhænge, der er mellem input og output. Det er nødvendigt med en god forståelse for at kunne vælge den tilgang og de algoritmer, der kan resultere i et brugbart system.

Der er to trin i at bruge supervised learning til at konstruere et system. Første trin er at klargøre data og vælge tilgang og algoritmer. Andet trin er læringsfasen, hvor de valgte algoritmer bruges til at generalisere fra eksemplerne. Resultatet er et system, der kan foreslå output givet nye input.

klassemærkerne: 02.1 (Biblioteksteknik og -administration) og 02.13 (Klassifikation).

Metadata og søgning

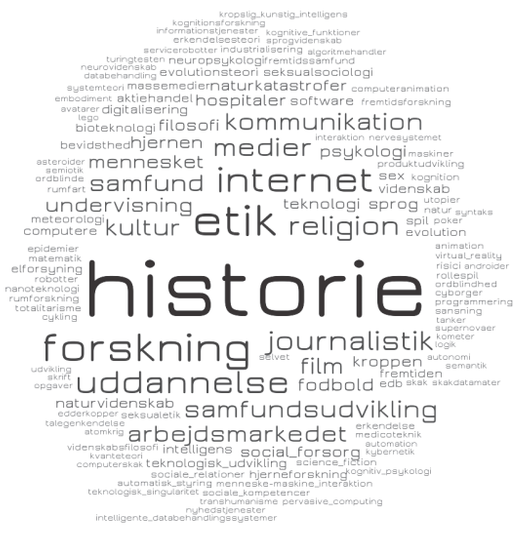
Formålet med metadata er at gøre det muligt for biblioteksbrugere at søge, og vi har da også kigget på, hvordan vi kan bruge analyse af metadata til direkte at forbedre søgemulighederne. Den del af vores arbejde har dog mere eksperimentel karakter.

Noget af det, vi har arbejdet med, er at skabe relationer imellem DBC's emneord. Disse er som udgangspunkt uden struktur, og man kan f.eks. ikke se, at emneordene 'hund' og 'gravhund' er relaterede, eller at 'historie' er et bredt begreb, mens 'poker' er et mere specifikt begreb.

Hvis der eksisterede relationer mellem emneordene, ville et søgesystem kunne bruge relationerne til at give brugerne bedre muligheder for at navigere i søgeresultaterne og søge videre. For eksempel kunne relationer mellem mere og mindre specifikke emneord bruges til at foreslå en bruger en bredere eller mere specifik søgning. Det første er især interessant, da det vil gøre det muligt at søge videre fra en søgning, der ikke gav resultater, fordi den var for specifik. Men det kunne også bruges til at give overblik over emneordene i et hierarki, som man kunne browse rundt i.

For at demonstrere mulighederne har vi i første omgang beregnet en primitiv 'bredde' for alle emneord, et tal, der angiver, hvor mange forskellige andre emneord emneordet forekommer sammen med.

Figur 1 er en måde at illustrere bredden af de forskellige emneord, der forekommer sammen med emneordet 'kunstig intelligens'. Stor skrift betyder, at det pågældende emneord bruges sammen med mange andre forskellige emneord, mens mindre skrift betyder, at det bruges sammen med færre andre emneord. En søgegrænseflade ville kunne vise resultatet af en søgning på 'kunstig intelligens' med en tilsvarende figur, og en bruger ville så kunne søge videre ved at vælge et eller flere emneord på figuren.



Figur 1. Visuel præsentation af emneord, der forekommer sammen med emneordet 'kunstig intelligens'. Skriftstørrelsen angiver, hvor mange andre forskellige emneord, et emneord bruges sammen med. Figuren er fra worditout.com.

Vi eksperimenterer også med at inddrage data om, hvilke materialer der er lånt sammen, og med at koble metadata sammen med udgangspunkt i det leksikalsk-semantiske ordnet DanNet, der udtrykker ords betydningsstruktur og interne relationer i

et formelt sprog, der gør dem anvendelige for it-systemer. (Se mere på wordnet.dk).

Ved at afdække relationer mellem metadata tror vi på store muligheder for at stille bedre søgemuligheder til rådighed, og det er derfor noget, vi vil arbejde videre med.

Mennesker og maskiner i fremtidens metadataskabelse

På baggrund af vores tværgående arbejde med metadataskabelse, statistik og machine learning tør vi allerede nu konkludere, at DBC i fremtiden vil kunne skabe mere værdi ud fra vores data ved at analysere dem systematisk.

Der er ikke noget, der tyder på, at maskiner kommer til at erstatte mennesker i arbejdet med metadata. Mennesket er stadig langt foran maskinen, når det gælder vores evne til at forstå indholdet i et materiale. Men vi tror på, at det engang i fremtiden vil være muligt at inddrage bibliografiske kompetencer på en ny måde – sammen med machine learning – så vi kan indeksere automatisk med meget høj kvalitet.