Statslige digitale publikationer – træningsdata

Statslige digitale publikationer indsamlet af Det Administrative Bibliotek. Datasættet er skabt af Det Kgl. Bibliotek, men defineret af Digitaliseringsstyrelsen med henblik på træning af danske sprogmodeller.

Beskrivelse af datasæt

Datasættet består af 14465 statslige digitale publikationer indsamlet af Det Administrative Bibliotek, der indtil 1. januar 2025 var Det Kgl. Biblioteks service for ansatte i centraladministrationen.

Datasættet er skabt af Det Kgl. Bibliotek, men defineret af Digitaliseringsstyrelsen med henblik på træning af danske sprogmodeller.

En del af teksterne er digitalt fødte tekster, mens en anden del af teksterne er digitaliseret med OCR-skanning. Ved tidspunktet for udgivelsen af datasættet har det været umuligt at skaffe oplysninger om, hvilke dele der er digitaliseret med OCR, men det antages, at materiale udgivet før 2000 er digitaliseret. Det har ikke været muligt at skaffe maskinproducerede OCR-konfidenstal til at beskrive OCR-kvaliteten, men kvaliteten af den digitaliserede tekst vurderes ved øjesyn at være god. OCR-kvaliteten er ikke forsøgt efterbehandlet med henblik på fejlrensning.

Hovedparten af teksterne er skrevet på dansk. En meget lille del er på forskellige andre sprog. Nogle ganske få tekster er skrevet på flere sprog. Det er f.eks. en vejledning på både dansk og grønlandsk.

Datasættet findes både i et csv-format og i parquetformat. Filerne er zippet med 7-Zip, som skal downloades og benyttes til at pakke filerne ud.

Databehandling

Oprindeligt datamateriale bestod af 16.009 XML-filer med struktureret information om publikationer og 15,965 txt-filer med tekstindhold fra publikationerne. Data er blevet filtreret for dubletter, filer mindre end 2 kb og tekster uden publiceringsår. På den måde er det endelig datasæt på 14465 tekster.

Til grundlag for databehandlingen ligger en datamodel, der er defineret af Digitaliseringsstyrelsen. Den beskriver kolonnenavne og værdier. På den baggrund er data fundet og trukket ud fra relevante felter i XML-filerne og behandlet således, at de stemmer overens med kolonnenavnene beskrevet i datamodellen. Behandlingen

inkluderer foruden matching af data også generering af nye data, f.eks. antal af ord, samt filtrering og rensning f.eks. af årstal.

Datasæt statistikker:

- 14551 rækker
- 17 kolonner
- 4.9GB filstørrelse csv-fil un-zippet
- 2.2GB filstørrelse parquet-fil un-zippet

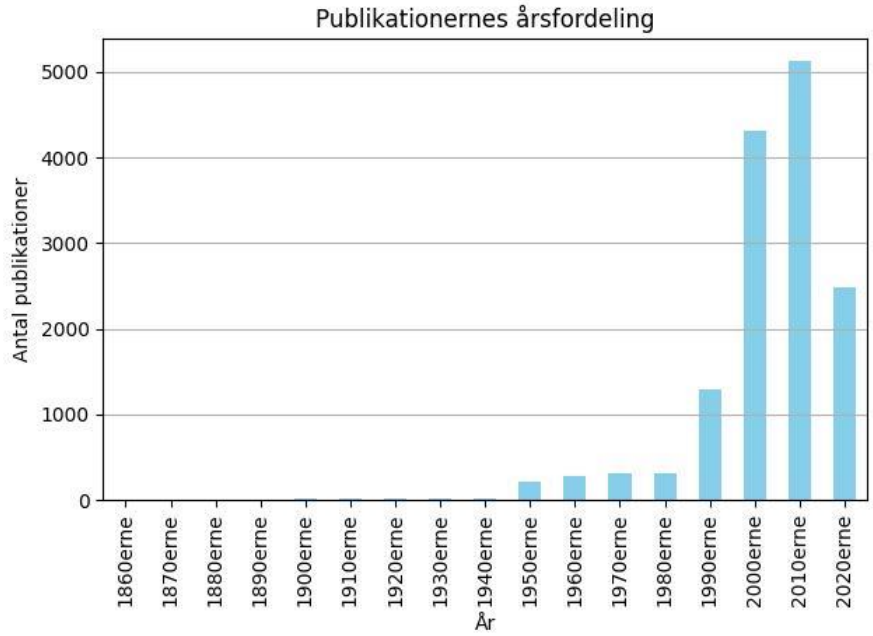
Beskrivelse af datasættets felter

Felt	Beskrivelse
Identifikator	En entydig reference til ressourcen. Det er ressourcens Persistent Identifier i bibliotekssystemet.
Indhold	Tekstudtræk fra den oprindelige kilde.
Skabt	Publikationens udgivelsesår.
Tilføjet	Dato hvor ressourcen er føjet til datasættet; dd-mm-åå.
Antal ord	Optælling af hvor mange ord ressourcen indeholder. Et ord opfattes som det, der står mellem to mellemrum.
Type	Type set i forhold til ressourcens genre eller iboende karakter. Det er Government publication (jf. https://id.loc.gov/vocabulary/marcgt.html)
Emne	Et centralt emne for ressourcen. En ressource kan have flere centrale emner. I dette tilfælde indeholder emnefeltet NaaN værdier, da der ikke er stringens i emnerne i de oprindelige metadata fra bibliotekssystemet.
Dannet ved OCR	En delmængde af datasættet er dannet ved OCR. Det har dog ikke været muligt at vurdere, hvor stor en delmængde.
OCR-metode	Det har ikke været muligt at fremskaffe information om ocr-metoden.

Fejlrenset OCR	Der er ikke fortaget efterbehandling for at fjerne OCR-fejl
Fejlrensningsmetode	Der er ikke foretaget forsøg på at forbedre OCR-kvaliteten.
Indeholder programmeringssprog	Datasættet indeholder ikke programmeringssprog eller opmærkning.
Efterbehandlet version af	Ressource er let renset, f.eks. er sekvenser af punktummer erstattet med et enkelt punktum.
Pseudonymiseret	Nej.
Pseudonymiseringsmetode	Ingen.
Sprog	Størstedelen af teksterne er på dansk. En lille del er skrevet på flere andre sprog og en lille del er på et andet sprog.
OCR- konfidensinterval	Der eksisterer i samlingen ikke information, der kan fortælle om konfidensintervallet.

Publikationer fordelt på år

Hovedparten af publikationerne er fra første del af 21. århundrede.



Licens

Love, administrative forskrifter, retsafgørelser, rapporter, betænkninger, vejledninger og lignende offentlige aktstykker er ikke genstand for ophavsret. Du kan frit kopiere, distribuere eller fremføre materialet uden at bede om tilladelse.

Persondata

Datasættet stammer fra statslige digitale publikationer offentliggjorte på ministeriers og styrelsers hjemmesider. Love, administrative forskrifter, retsafgørelser og lignende offentligt tilgængelige aktstykker er ikke undergivet særlige beskyttelseshensyn.

Du kan læse mere om Det Kgl. Biblioteks behandling af personoplysninger i de statslige digitale publikationer **her.**