Danmarks Breve - træningsdata

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

Beskrivelse af datasæt

Datasættet består af 13516 breve skrevet fra 1500-tallet til 1900-tallet af danskere med historisk og kulturel betydning. Det er skabt af Det Kgl. Bibliotek, men er defineret af Digitaliseringsstyrelsen med henblik på træning af danske sprogmodeller.

Datasættets tekst indhold, samt metadata baserer sig på de TEI-filer, der bliver benyttet til Det Kgl. Biblioteks samlingen "Danmarks Breve". Samlingen indeholder digitaliseringer af en lang række trykte brevudgivelser fra Det Kgl. Biblioteks samlinger, udgivet over en længere årrække af mange forskellige udgivere.

Teksterne er digitaliseret med OCR-skanning. 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.

Størstedelen af tekstindholdet er dansk og hovedsageligt historisk dansk. En mindre del af indholdet er på et andet sprog, f.eks. tysk og engelsk.

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

Databehandling

Oprindeligt datamateriale bestod af lidt mere end 13000 TEI-filer dels med relativ velstruktureret metadata om brevene, dels med brevteksten.

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:

- 13516 rækker
- 17 kolonner
- 88MB filstørrelse csv-fil un-zippet
- 53MB filstørrelse parquet-fil un-zippet

Beskrivelse af datasættets felter

Felt	Beskrivelse
Identifikator	En reference til den digitale kilde i Det Kgl. Biblioteks digitale samlinger.
Indhold	Brevtekst.
Skabt	Årstal fra brevets dato eller datering.
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	Brev.
Emne	Ikke brugt
Dannet ved OCR	Tekstindholdet er dannet med OCR skanning.
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	En lettere efterbehandlet version af teksten i kolonnen indhold.

Pseudonymiseret	Nej.
Pseudonymiseringsmetode	Ingen.
Sprog	Størstedelen af tekstindholdet er dansk og hovedsageligt historisk dansk. En mindre del af indholdet er på et andet sprog, f.eks. tysk og engelsk.
OCR- konfidensinterval	Der eksisterer i samlingen ikke information, der kan fortælle om konfidensintervallet.

Licens