Digitaliserings-
styrelsen

18. marts 2024

Dokumentation for scrape af Region H dokumentssamling

Information om korpus

Korpusset indeholder alle Region Hovedstadens vejledninger, instrukser, og politiker, som ikke er login-beskyttede, hentet fra vip.regionh.dk. Dokumenterne er gemt som txt-filer med UTF8-kodning.

Datasættet må ikke anvendes som sundhedsfaglig informationskilde:

Der gøres opmærksom på, at korpusset er udviklet med henblik på udvikling af sprogteknologi og må ikke bruges som sundhedsfaglig informationskilde. Dokumenterne er scrapet på et specifikt tidspunkt og vil derfor ikke være ajourført. Der henvises i den forbindelse til Region Hovedstadens dokumentsamling

Filnavne

Filerne er navngivet efter udgiver og titel på dokumentet. Eksempelvis er dokumentet med titlen "B-Erytroblaster", som er udgivet af Klinisk Biokemisk Afdeling på Nordsjællands Hospital, navngivet "Nordsjællands Hospital_Klinisk Biokemisk Afdeling_B-Erytroblaster". I nogle tilfælde er udgiver og/eller titel forkortet. Mapping mellem filnavn og udgiver og titel findes i "Filnavne.xlsx".

Tabeller

Tabeller er fjernet fra dokumenterne og gemt i separate filer med samme filnavn, men med tilføjelsen "_TABEL_X", hvor X er tabellens nummer i dokumentet startende fra 1. I dokumentet er tabellerne erstattet med tagget "###TABEL_X###". Antallet af tabeller per dokument findes i "Filnavne.xlsx".

Subkorpusser

Korpusset indeholder to subkorpusser: Et råt korpus og et renset korpus. Tabel 1 indeholder information om de to subkorpusser.

	Råt korpus	Renset korpus
Filer	24.752	24.752
- Dokumenter	15.829	15.829

- Tabeller	8.923	8.923
Tokens	12.232.517	9.941.236
Chars	364.971.020	79.942.659

Tabel 1: Information om subkorpusser.

Råt korpus

Det rå korpus indeholder de rå dokumenter med alle opmærknings- og programmeringssprog startende med dokumentets titel og indtil teksten slutter.
Bemærk at JavaScript egenskaben ”display: none;” er brugt i nogle af dokumenterne, således at hvis teksten fremvises i en HTML-viewer, vil den være skjult som udgangspunkt.

Renset korpus

Det rensede korpus er renset for alle opmærknings- og programmeringssprog og opstillet i et menneskeligt læsbart format. Dette inkluderer:

- Håndtering af unicode chars som angivet i Tabel 2.
- HTML char referencer som '<' (’<’) og '&' (’&’) er erstattet af deres tilsvarende char.
- Standardfraser uden betydning i teksten er slettet, for eksempel ”Tilbage til top”.
- HTML lister er erstattet med et indryk med ”- ”. I tilfælde af lister med sublister, er der ikke lavet yderligere indryk.
- Der er linjeskift efter en overskrift.
- Alle overskrifter undtagen dokumentets titel i linje 1 efterfølger to linjeskift.
- Mellemrum og linjeskift er strippet til et enkelt bortset fra i ovenstående tilfælde.
- Tabeller er præsenteret celle for celle adskilt af linjeskift, startende med kolonne 1 til n i række 1, derefter række 2, osv.

Anonymisering

Vær opmærksom på at det rensede korpus ikke indeholder alle anonymisering af for eksempel foldere, da de kan være del af en struktur, som er blevet slettet i forbindelse med rensningen.

Filer indeholdende navne blev udvalgt til manuel anonymisering ved brug af en svagt superviseret deep learning model trænet ud fra metoden beskrevet i [1]. Filer indeholdende resten af typerne blev udvalgt ved brug af regular expression søgemønstre. Desuden havde annotatorerne mulighed for at kategorisere tekst i en diverse kategori ved manuel gennemgang af dokumenterne, hvis en anonymisering faldt uden for de prædefinerede grupper.

Unicode kategori	Håndtering
Control	Slettes, undtagen: u'\u000a' (‘\n’) slettes ikke u'\u0009' erstattes af ’ ’
Format	Slettes
Private use	Slettes
Space separator	Erstattes af ‘ ‘, undtagen: u’\u0020’ (‘ ‘) slettes ikke
Paragraph separator	Erstattes af ‘\n’
Line separator	Erstattes af ‘\n’

Tabel 2: Håndtering af unicode chars

Anonymisering

6.028 filer, svarende til 24% af korpusset, blev udtaget til manuel anonymisering. Tabel 3 viser de anonymiserede oplysninger, det tag som de er erstattet med i teksten, og antallet af anonymiseringer fordelt på type.

Type	Tag	Antal
Foldere på server	###FOLDER###	31.611
Telefonnumre	###TELEFON###	6.701
E-mailadresser	###EMAIL###	829
Navne på personer	###NAVN###	5.737
IP- og MAC-adresser	###IP_MAC###	10
CPR-numre	###CPR###	8
Personadresser	###ADRESSE###	37
Diverse	###DIVERSE###	110

Tabel 3: Overblik over anonymisering

Laursen, M. S., Pedersen, J. S., Vinholt, P. J., & Savarimuthu, T. R. (2023). Automatic Annotation of Training Data for Deep Learning Based De-identification of Narrative Clinical Text.