

Refleksion

Portfolio 1's udbytte tager grundlæggende basis i en forståelse af et datasæt og hvordan der arbejdes med det. Derved er udbyttet af denne opgave essentielt, fordi ved dette blev det gjort muligt, at viderebygge på denne grundforståelse og arbejde med mere komplicerede opgaver inden for åbne datasæt.

Grundforståelsen vil være anvendelig i fremtidige datasæt, hvor opgaver som at kunne vise antal *rows*, *datatyper* og hvordan statistiske data trækkes ud af datasættet, vil være brugbart.

Portfolio 2 bygger videre på denne grundforståelse, med arbejdet af TheGuardian, der var en *open api*, hvor det nu var et større datasæt og der var mulighed for afgrænsning i form af, hvilken periode dataet skulle være fra. Dernæst gav den en større forståelse af, hvordan et datasæt afgrænses, da opgaver som at finde *tf.idf weight*, *stopwords*, *tokenization*, *unique words* osv., inden for et specifikt emne, gjorde, at det nu var muligt, at finde relevant materiale inden for et stort datasæt. Udbyttet ved dette består i praksis af, at det nu er muligt, at arbejde med datasæt og det vil være muligt, at anvende ude fra studierne på aktuelle datasæt for institutioner, der har brug for, at få behandlet deres data.

Portfolio 3's udbytte består i forståelsen af, hvordan en URL kan konstrueres og dets praksis kan være anvendelig ude fra studierelaterede forhold, hvor oprettelsen af en URL forekommer ofte. Ydermere, hvilke metadata, der kan blive anvendt via Dublin Core skabelonen, hvor dette kan være udgangspunktet, hvis en API eller et datasæt skal konstrueres. Hvordan uddrag af datasættet kan visualiseres på forskellige måder via diagrammer og skabeloner, var også en af de pågældende opgaver og dette kan anvendes ved præsentationer af dataet både i og ude fra studierelaterede forhold.

Etiske overvejelser

Ved Titanic datasættet og SMK API'en blev der arbejdet med navne, alder, nationalitet, fødested osv.

Derfor kan dette anskues som personfølsomme data og skal derfor behandles med respekt. Resultaterne ud fra maskinlæring via Titanic datasættet og SMK API'en, kan derved ud fra Wilkinson et al. (2016) FAIR princip, skulle gøres let tilgængeligt og anvendelig. Ved dette kan det være problematisk, da personerne det omhandler eller dem, der har relationer til informationerne, ikke føler, de bør blive delt med andre. Der skal derfor som Hampton et al. (2015) beskriver "Likewise, ecologists may be reluctant to share data that could harm people (e.g., by lowering property values)".

Der kan argumenteres for, at personfølsomme data kan skade andre, som Hampton et al. (2015) beskriver, og det skal derfor vægtes mod Wilkinson et al. (2016) FAIR princip, da resultater fundet ud fra disse

datasæt også skal kunne deles med andre, om det kan være anvendeligt, og på denne måde kunne fremme udvikling og forskning.

Dette skal også opretholdes mod arbejdet, der er udført med de førnævnte datasæt, og være en overvejelse. Men ydermere med anden forskning, der kan indeholde personfølsom data, hvor det kan være direkte skadeligt for nogle parter, skal en grundig overvejelse være tilstede om, hvordan disse resultater gøres tilgængeligt, både for offentligheden og andre forskere. Da resultaterne kan være med til at fremme forskning og udvikling, skal der derfor heller ikke være en "I own the data" fornemmelse som Hampton et al. (2015) beskriver, fordi dette heller ikke vil gavne omgivelserne, hvis resultaterne ikke bliver taget i brug. Det er derfor en gylden mellemvej om, hvor meget det gavner eller kan skade, der skal tages i forbehold og der kan stærkt argumenteres for, at ved den mindste risiko for det kan skade andre, skal det ikke offentliggøres.

Litteraturliste

Hampton, SE, et al. (2015) The Tao of open science for ecology. *Ecosphere* 6(7) article 120.
<https://doi.org/10.1890/ES14-00402.1>

Wilkinson, M. D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 <https://doi.org/10.1038/sdata.2016.18>