Den samlede portfolio: refleksioner og etiske overvejelser (reflections)

Faget Open Data Science har omhandlet arbejdet med åbne data og analyse af hvordan der kan arbejdes med denne data i praksis. De tre portfolioer har alle medvirket til en forståelse af hvordan jeg kan arbejde med åbne data, om det er til eget brug eller arbejdsrelaterede opgaver. Dette dokument vil, som overskriften antyder, reflektere over arbejdet med åbne data ud fra de tre portfolioer i form af læringsbytte og fremtidig brug af det lærte. Dette vil deles op efter de tre portfolioer. Derudover vil der også argumenteres for de etiske overvejelser jeg har haft undervejs, samt dem der kan vendes efter dette fags ende.

Portfolio 1 var den første vi fik stillet, og den arbejdede ud fra de øvelser og viden vi havde fået om kodningssproget Python og bl.a. Pandas biblioteket. Her lærte jeg hvordan jeg kunne udtrække data fra en csv fil om Titanic og dets passagerer. Jeg fandt det yderst interessant hvordan jeg kunne bruge data, især tallene fra CSV-filen, til at lave en pivot tabel. Det er noget jeg før har kendt fra programmer som Microsoft Excel, så det at lære at kunne gøre det i samme program, hvor man koder var meget læringsrigt.

The second portfolio assignment consisted of demonstrating how to use text mining work on a collection of data. In this case I had the opportunity of choosing between two kinds of APIs' and I went with the Guardian API. Here I got to work further with the smaller exercises we had during the course into a larger scale. For topic modeling I chose the term 'Scotland' and made it my query. This resulted in me discovering the most used words in articles with this query in it, which I also had a word cloud show. All in all, this portfolio gave me a great view into how to use text mining on data, and how, both in personal but also academic life, to analyze collected data.

Den sidste portfolio omhandlede også en API, denne gang SMKs beta version af den. Jeg fandt det interessant at vælge kunstneren Poul S. Christiansen, da den åbne samling online viste, at SMK havde flere af hans malerier og tegninger. At få API'en til at virke gav dog nogle problemer, da søgningen ikke var så ligetil, og kun en bestemt måde at skrive kunsternes navn på gav resultater. Da der kom en god url ud med data i JSON-format blev det arbejdet med i Jupyter Notebook, hvilket også er et godt program til at kode i. Selve opgaven var virkelig spændende, og har også givet mig lysten til at arbejde videre med API'er med kunstværkers data. Et af de mest interessante punkter i denne portfolio var brugen af altair biblioteket, der kan vise flotte grafer. Til

min besvarelse valgte jeg bl.a. en color-graph, der fremviste hvordan farverne på malerierne i API'en stemte overens med det man så på SMK's hjemmesides samling af værker.
Dataetik er en vigtig del af at arbejde med åbne data, samt overvejelserne bagom. Åbent data er blevet mere og mere tilgængelig, men det er vigtigt at forstå hvilken slags data man har med at gøre, samt hvilke regler der ligger bag:

> Open Science is about increased rigour, accountability, and reproducibility for research. It is based on the principles of inclusion, fairness, equity, and sharing, and ultimately seeks to change the way research is done, who is involved and how it is valued. It aims to make research more open to participation, review/refutation, improvement and (re)use for the world to benefit (Bezjak et al., 2018)

Som citatet ovenfor fortæller, så handler åbne data om at samle folk i at kunne bruge data. Førhen har data været lukket væk kun for forskere og folk inden for det akademiske, men nu med åbne data forsøges der at ændre hvilke personer kan forske med data (Bezjak, 2018). Men sådan en proces kan også skabe forvirring omkring forandring, om hvad det indebærer. Hampton et al. (2015) klargør den stigma, der må forekomme, når forskere pludselig skal åbne data forskningsproces mere op og tidligere end de har gjort før. I visse tillfælde kom det frem, at nogle forskere holdte deres forskning hemmelig før publicering:"*(…) to avoid the risk of looking foolish*" (Hampton et al., 2015, s. 7). Det kan godt anses for at være godt at dele processen mere for at skabe rum for forbedring og god dialog med andre forskere (Hampton et al., 2015, s. 10).

En anden etisk overvejelse med arbejdet af åbne data er personerne i dataet. Som navnet indebærer, skal åbne data jo være tilgængeligt for alle, og også kunne bruges til arbejde, men et dilemma opstår indenfor privatlivet og persondata. ACM (Association for Computing Machinery) skildrer det løfte, der skal afholdes omkring datasikkerhed overfor privatlivet, hvor der skal tages:" *(…) precautions to prevent re-identification of anonymized data or unauthorized data collection, ensuring the accuracy of data, understanding the provenance of the data, and protecting it from unauthorized access and accidental disclosure*" ("ACM Code of Ethics and Professional Conduct", (n.d.)). Dette tydeliggør hvordan åbne data skal være privat på det punkt, at identificerbare data om privatpersoner ikke skal kunne afspores eller deles. Dette giver også nogle klare rammer om

hvad åbne data skal være. Derudover kan spørgsmålet omkring termet *machine learning*, og hvad det indebærer også nævnes. Dette omhandler brugen af computere der skal kunne læse og lære af den data den gives. Dette kan både have fordele og ulemper. Fordelen er effektiviteten og den hurtige arbejdspotentiale det giver i form af f.eks. opgaver med filtrering af filer, men i områder med persondata kan det ses vigtigt med personer involveret i arbejdet. Som ACM viste, så findes der flere regler indenfor det etiske med arbejdet med databehandling, hvor private og fortrolige oplysninger skal omgås med nøje varsomhed og respekt. Dette kan ikke forstås på samme måde med computere, der er programmeret til at lagre og arbejde på en bestemt måde, og her kan især følsomme informationer tænkes som sårbare i form af f.eks. anonymitet og identificerbarhed til de enkelte individer (”ACM Code of Ethics and Professional Conduct”, (n.d.)). Dette er et vigtigt punkt at have med når der arbejdes med data og *machine learning*.

Det kan konkluderes at åbne data er fremtrædende indenfor forskning, også for privatpersoner eller studerende, men at det samtidig skal overholde dataetiske regler, og kunne forsikre arbejde med data, der ikke overtræder dataetiske regler.

Litteraturliste

ACM Code of Ethics and Professional Conduct. (n.d.).
Retrieved from https://www.acm.org/code-of-ethics

Hampton, S. E., Anderson, S. S., Bagby, S. C., Gries, C., Han, X., Hart, E. M., … Zimmerman, N. (2015). The Tao of open science for ecology. *Ecosphere*, *6*(7). doi: 10.1890/es14-00402.1

Sonja Bezjak, April Clyburne-Sherin, Philipp Conzett, Pedro Fernandes, Edit Görögh, Kerstin Helbig, … Lambert Heller. (2018). Open Science Training Handbook (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.1212496