Etik

Facebook/Cambridge Analytics skandalen, Yahoo skandalen, PlayStation Hack osv. Alle disse har med store datamængder og kompromittering af især persondata. Brugernes personlige data er endt hos personer, der har haft dårlige intentioner med det. Dette har bl.a. medført EU´s persondataforordning, som stiller krav til virksomheder omkring det indsamlet data. Foruden denne lov findes der heldigvis mange muligheder og retningslinjer for, hvordan man kan arbejde med data science uden at komme i karambolage med lovgivning eller ophavsrettighedshaveren. ACM Code of Ethics and Professional Conduct, er et godt eksempel på dette. Hvor organisationen arbejder med guidelines for, hvordan man kan arbejde med data på et etisk plan.

"The Code is designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. Additionally, the Code serves as a basis for remediation when violations occur." ACM (2020) (https://www.acm.org/code-of-ethics).

Til sammenligning i faget Open Data Science har der ikke været mange datasæt, der har indeholdt personfølsomme data. Titanic datasættet, brugt i portfolio 1 opgaven, havde alder og navne på passagerne. Denne information må siges at være offentlig kendt siden 1912, og overlevende personer er for længst gået bort. Portfolio 2 opgaven havde avisen The Guardian som datasæt, og derved var alle de tekstfiler vi arbejdede med, offentlig tilgængelige avisartikler. Tredje portfolio havde Statens Museum for Kunst som datasæt og her var mange data, der kunne være personfølsomme. En aktuel kunstner, såsom Olafur Eliasson, er at finde i datasættet, men denne viden er ikke noget man nemt kan finde på f.eks. Wikipedia.

I forsknings og akademiske sammenhænge er det vigtigt at have det etiske aspekt, omhandlende overordnet hvad denne data skal bruges til. Er det nødvendigt for den pågældende forskning at have al persondata, eller kan man nøjes med en simpel person beskyttet udgave? Det kunne tænkes, at der blot blev indsamlet alder og køn, og ikke navn til et forskningsprojekt, fordi navnet eller andet persondata alligevel skulle anonymiseres.

Refleksion

Faget Open data science, har givet mig et højt og interessant læringsudbytte. Niveauet for dette fag har været ekstremt højt, til tider måske for højt, men dette har presset mig til at lære og undersøge yderligere om bl.a. grundkodning og funktioner i Python, ved siden af faget. Især modulet med Frans har været givtigt, da jeg føler, at jeg har lært metoder til hvordan store mængder af data kan

behandles og analyseres. I denne sammenhæng har tf-idf (term frequency-inverde document frequency) været det mest lærende aspekt. Kurset har også givet frustrationer i form af manglende viden, hvor jeg er endt med at sidde til langt ud på natten, og googlet mig til diverse svar på stackoverflow og lignende sider.

Flere af de tillærte funktioner har jeg allerede benyttet mig af i mit bachelorprojekt, og i fremtiden kunne det være interessant, at benytte nogle af disse tilegnede færdigheder til at analysere kæmpe datamængder, for at kunne gøre søge termer og muligheder bedre for slutbrugeren. Her kunne tænkes, at man kunne lave et projekt med et bibliotek eller en større virksomhed, for at analysere på deres data, for derved at lave f.eks. topic models og hurtigere kunne finde frem til emner ~~hurtigere~~ på denne måde, end at bruge gængse søgemetoder. En af de vigtigste konklusioner, som jeg har fået på dette fag, omhandler den tilgang, hvorved et nyt datasæt skal oprettes. Her er det vigtigt at have et samlet regelsæt for hvilken data der skal med, og ikke mindst det data der ikke skal med. Her kan f.eks. Dublin Core (DMCI 2020) ~~her~~ være en god tilgang og retningslinje for hvilke data, der skal medtages i et nyt sæt. Et andet metadata termsæt, som jeg har fundet mere brugbart, har været Schema.org (2020). Dette er udviklet af Google, Microsoft, Yahoo og Yandex, til netop at kunne fungere udelukkende online og på tværs af systemer og platforme. Denne liste af termer er virkelig lang i forhold til Dublin Core, men rammer bredt. Mange ting kan undlades, men det smarte ved denne er underkategorierne, som kan give flere muligheder, et eksempel er "DayOfWeek", som har de syv ugedage selvfølgelig, men også medregner "PublicHolidays", som kan falde alle ugedage, men f.eks. falder juleaften i Danmark altid d. 24/12 og i Storbritannien d. 25/12, derfor skifter den ikke bare fra år til år, men og kulturelt. Denne form, med mange metadata termer, gør det nemmere at læse for systemer mellem hinanden, end Dublin Core, hvor få termer kan få skaberen af datasættet til at lave termer der passer til eget datasæt, f.eks. "dateAccepted" og "dateSubmitted" i DCMI, ville formentlig ikke være relevant i en database over film, og "date" er ikke rammende, når der tænkes udgivelsesdato.

Kommenterede [CRH1]: Tror det skal skrives på en anden måde?

Kommenterede [CRH2]: Læse for?

Som supplement og bedre forståelse til dette fag, har jeg i semestret taget kursus i NVivo, som kan nogle simple ting, såsom Word clouds, og simpel tekstanalyse. Dette valgte jeg, for at kunne have en bedre forståelse af hvordan visse programmer virker, som kan det samme, som vi har lært i faget, men i mindre skala end f.eks. Python.