# Open Data Science – Portfolio eksamensaflevering

**Eget læringsudbytte**

Jeg havde allerede et godt, men primært teoretisk, kendskab om Open Science qua mit arbejde på et universitetsbibliotek og min position som koordinator for AU Librarys indsats på området. Min motivation til at melde mig til kurset var at få en dybere forståelse for, hvordan man rent praktisk arbejder med data via API, men også at arbejde seriøst med programmeringssproget python, analysere data og arbejde med konkrete projekter.

Mine forventninger til læringsudbyttet er blevet opfyldt på den måde, at jeg ved kursets afslutning har fået kendskab til en række API’er, er blevet trænet i at arbejde i miljøer som Spyder og Jupyter Notebooks, og at jeg har en meget større indsigt i de forskellige muligheder som pakkerne Pandas og Sklearn og Numpy giver i forhold til visualisering og analyse af store datamængder. Det giver et rigtigt godt afsæt til i sær to ting: For det første at gå videre med kodning, og få en bedre grundlæggende forståelse for kodesproget og logikken bag, og for det andet til at forstå hvad der sker bag brugergrænsefladen på redskaber vi allerede formidler, som NVivo og Voyant.

Rent praktisk vil den opnåede viden udmønte sig i at arbejde med at give både den teoretiske og praktiske viden videre til kolleger på Det Kgl. Bibliotek, inspirere studerende i datalabs og kunne give forskere en større viden om bibliotekets kulturarvsdata. Selv om man ikke bliver professionel koder på 3 måneder, vil dette kursus alt i alt være et rigtigt godt afsæt til mere kompetenceudvikling indenfor den praktiske side af datascience, og til at arbejde videre mere at formidle kulturarvs- og andre åbne datasæt.

**Etiske overvejelser i data science**

Overordnet set ligger arbejdet med store mængder data under de samme etiske retningsliner, som alt andet forskning, nemlig den danske Code Of Conduct[[1]](#footnote-1), som alle universiteter har skrevet under på at de vil følge[[2]](#footnote-2). De tre hovedmantraer er ærlighed, gennemsigtighed og ansvar. Ærlighed i forhold til rapportering af resultater, gennemsigtighed i forhold til metode og konklusioner og ansvar for at alle regler er overholdt, herunder de etiske retningsliner. Man skal kun tilbage til foråret for at se hvor galt det kan gå, når de etiske retningslinjer ikke bliver overholdt, som i sagen om oksekødsrapporten, hvor forskere fra Nationalt Center for Fødevarer og Jordbrug (DCA) på Aarhus Universitet havde indgået et lidt for tæt samarbejde med interesseorganisationen Landbrug og fødevarer og Danish Crown, og rapporten ikke afspejlede samarbejdets natur og indflydelse på resultaterne[[3]](#footnote-3).

Når det gælder Open Science er der endnu et etisk imperativ, der handler om at dele, og mottoet er ”As open as possible, as closed as necsseary”, hvilket man bl.a. kan finde helt tilbage til Open Access bevægelsens opståen. Open Acces blev en grundsten for at søge forskningsmidler under Horizon 2020, EU's pulje til forskningsprojekter[[4]](#footnote-4). Ønsket om Open Acces til publikationer er senere blevet udvidet til at omfatte også data[[5]](#footnote-5), og FAIR principperne[[6]](#footnote-6) fra 2016 omfatter forskningsdata fra hele workflowet, og målet er at gøre data Findable, Accesible, Interoperable og Reusable. Det vil sige ikke bare give adgang, men at give adgang på en måde, så formatet er anvendeligt, at metadata er forklaret på en måde, så man forstår hvad data består af, men også provenance, altså datas historie i forhold til indsamling, sortering, og så videre. At FARIficere sine data betyder, at man gør dem så åbne som muligt, eller as open as possible.

Men den åbne politik udløser et endnu større ansvar i forhold til den anden side af mønten: ”As closed as necessary”. Især efter den skærpede databeskyttelseslov endeligt blev effekturet i 2018[[7]](#footnote-7) er opmærksomheden på dels personhenførbare data og dels særlige følsomme personoplysninger blevet øget, ligesom folks ret til egne data og mulighed for at få slettet disse i offentlige systemer.

Ser man på den data vi har arbejdet med i kurset, er det åbne data med lav følsomhed, og primært historiske data. Selv om data fra portfolie 1 er personoverførbare fordi det fulde navn indgår, er data over 100 år gamle, og det kan i den sammenhæng være svært at se nogle særlige etiske problemstillinger i forhold til det datasæt. SMKs åbne API der kalder metadata fra kunstværker, er heller ikke problematiske i forhold til overnævnte problemstillinger, men der kan være andre overvejelser i forhold til hvilke data man indfører. Skriver man noter ind, der indeholder ord, der er stødende for nogle dele af samfundet i dag, som fx neger? Og hvordan håndterer man værker, der er anskaffet på en måde vi ikke vil mene er etisk korrekt i dag, som fx fra diktaturer?

Ser man på nyere data som The Guadians API, kan der være andre problemstillinger. Selv om data er åbent, om allerede udgivet vil en ny samstilling og udgivelse af data kunne være problematisk i forhold til ACM’s etiske regelsæt[[8]](#footnote-8), hvor punkt 1.4 lyder ”Be fair and take action not to discriminate.”. Laver man en topic modeling analyse, på et datasæt, der allerede er vægtet i forhold til negative holdninger overfor indvandrere uden at gøre det fuldstændigt gennemsigt og man ærligt redegør for hvordan man har udvalgt corpus, vil det give et forvrænget billede af debatten, der ikke lever op til hverken ACM’s etiske regelsæt, men heller ikke den danske code of conduct, der altså lægger vægt på ærlighed i rapportering og gennemsigtighed overfor metode.

**Litteratur**

Drivsholm, Bahn & Andersen (2019, 30.november) Da landbrugslobbyen ville ændre oksekøds dårlige omdømme: Sådan gik det galt på Aarhus Universitet. *Information, Videnskab.s.12*

# Portfolio 1 - tekst

For at kunne bestemme hvilke resultater der kunne være interessant at trække ud af dette datasæt fra Titanic, er der set på datasættet i sin helhed. Data består af deskriptive data som køn, alder, og hvilken klasse passagererne er rejst på. De beskriver lidt over en fjerdedel af del samlede antal passagerer.

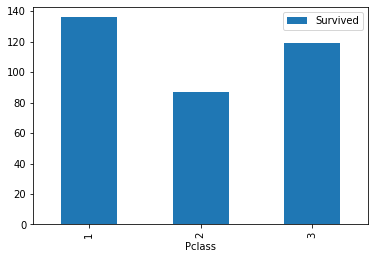
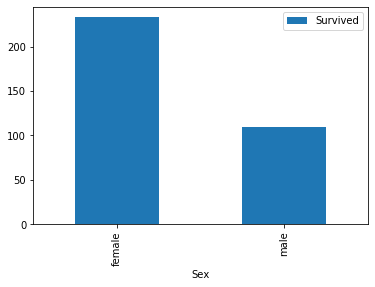
Ovenstående data er let aflæselige, men der er også en kategori som medrejsende søskende/forældre og børn/ægtefælle som er lidt sværere at have med at have med at gøre med i forhold til at trække nogle meningsfulde statistikker ud. Prisen på rejsen ser også ud til at svinge voldsomt fra passager til passager, så uden yderligere information om, hvordan de er udregnet, er det også en kategori, der er svær at have med at gøre analytisk.

Dernæst er der kigget på hvilke slags data, som data sættet består af, for at se på hvordan man kan bruge hvilke kolonner, og hvordan de kan behandles, ved hjælp af en simpel vise kommando. Sættet der hentes ind som en dataframe består af tal som intergers(hele tal), floats(kommatal) og strings som objects.

Derefter er der brugt simple funktioner, der ligger i pandas for at kunne besvare spørgsmål om størrelsen på datasættet, spørgsmål om den yngste og ældste passager samt middelværdi af alder på de 887 passagerer i sættet.

For at besvare spørgsmålet om der ens efternavne i sættet, er der først forsøgt at isolere efternavne ved at splitte objektet navn, for derefter at køre en funktion, der viser om der er duplikater i sættet. Analysen viser, at der findes duplikater, men da objekterne var uens af længde, og det derfor var vanskeligt at isolere efternavnet, er analysen ikke helt korrekt udført.

Til sidst viser en pivot tabel forskellen på antal overlevende passagerer på de tre klasser. For at visualisere tabellen og gøre resultatet mere umiddelbart af aflæse, er tabellen herefter lavet til en barchart med funktionen plot.bar():

￼￼￼￼￼￼

￼￼

Og til allersidst er der lavet en visning af antal kvinder kontra mænd der overlevede.

Konklusion ￼￼

De tre små analyser til sidst viser lidt om hvad der er muligt at foretage sig i et python-script med ganske lidt øvelse. Skulle man arbejde videre med tallene ville man skulle se lidt mere på hvilke analyser der ville give mening statistisk set. Antal kvinder og mænd i den sidste analyse ville for eksempel med fordel kunne regnes ud til procent.

Datasættet i sig selv har også nogle svagheder med uigennemsigtige tal og uklare definitioner.

**Kode**

|  |
| --- |
|  |
|  | #a) henter dataset |
|  | import pandas as pd |
|  |  |
|  | import numpy as np |
|  | pd.options.display.max\_rows = 900 |
|  |  |
|  | titanic=pd.read\_csv('titanic.csv') |
|  |  |
|  | #b)#viser at der ikke mangler data |
|  | print(titanic.isnull().sum()) |
|  |  |
|  |  |
|  | #opgave 3) beskriver datasettet |
|  |  |
|  | titanic.shape |
|  | titanic.size |
|  | titanic.dtypes |
|  | print('Datasettet Titanic har antal columns og rækker ',titanic.shape, 'og',titanic.size, 'antal værdier.') |
|  | print('Og der er disse typer data',titanic.dtypes) |
|  | print('Middelværdi alder er ',titanic['Age'].mean()) |
|  | print('Yngste rejsende var',titanic['Age'].min(),'år gammel') |
|  | print('Ældste rejsende var',titanic['Age'].max(),'år gammel') |
|  |  |
|  | #antal døde |
|  | number\_travelers = 887 |
|  | number\_survivers = 342 |
|  | number\_dead = (number\_travelers - number\_survivers) |
|  | print('Der var ', number\_dead,'mænd, kvinder og børn, der omkom på rejsen' ) |
|  |  |
|  | #Opgave 4 - viser at der ens efternavne i datasettet. |
|  | Names = (titanic['Name'].str.split(' ', n = 2, expand = True)) |
|  | first\_Names = (Names[1]) |
|  | last\_Names = (Names [2]) |
|  | last\_Names.duplicated(keep='first') |
|  |  |
|  | #Opgave 5 pivot tabel |
|  | #pivot tabel der viser antal overlevende på hhv første anden og tredie klasse. |
|  | pivot=pd.pivot\_table(titanic,'Survived', ['Pclass']) |
|  |  |
|  | print (pivot) |
|  |  |
|  | # skaber en pivotabel vist i barchart med antal overlevende på de tre klasser. |
|  | pivot=pd.pivot\_table(titanic,'Survived', ['Pclass'],aggfunc= np.sum).plot.bar() |
|  |  |
|  | print (pivot) |
|  |  |
|  | # skaber en pivotabel vist i barchart med antal overlevende på de tre klasser. |
|  | pivot=pd.pivot\_table(titanic,'Survived', ['Sex'],aggfunc= np.sum).plot.bar() |
|  |  |
|  | print (pivot) |
|  |  |
|  |  |

# Portfolio 2

**The Guardian API**

The British newspaper The Guardian was first released in 1821, and is now published in the UK, the US and in Australia. The Guardian Open Platform gives free access to almost two million items of content, access to full text, and has a very user friendly and developed documentation. This portfolio is working with the Guardian API, and the results are from all branches of the guardian, and all sections.

**Importing the text**

The text is imported by using a script found on github[[9]](#footnote-9) that collects all content between to dates, set as parameters for the API request. It continues to make a list of texts. This is the corpus used in the analysis.

Looking at the data

To get a feel of the text some simple functions is called. To find out the size of the corpus, the number of texts is called. To have a closer look, the random.sample function is called on full text. This gives a view of two full texts articles. The number of words is also counted by calling the word tokenize function in the nltk library, and then making a list of all words, and the random function called again. This is all to check the import and see if the corpus makes sense.

Number of words and sentences is used to calculate the average number of words and sentences per article. This gives a sense of the types of texts the corpus is comprised of. If these are articles, they are fairly long reads.

Term document matrix

Using a term document matrix allows us to make a matrix with an index for texts, and index for words, and a grid with values for every word used in the texts. The stop words are removed in this step, as well as normalizing the vocabulary with a token pattern.

SKlearns count vectorizer is imported, and the count vectorizer is fitted to the vocabulary of the texts, and then the transform function is called to encode the texts to the vector.

This matrix has a very hi sparsity which means most of its values are zero. The matrix allows us to find the top words used after cleaning the data.

Applying TF-IDF weighting

The term frequency-inverse document frequency matrix gives the weight of the words or the relevance. This means that doing the same calculation on the new TF-IDF matrix, the results are much more rare words with more meaning, as the new matrix not just counts the frequency words, but sets it against how common they are in the text corpus. This gives us the relevance for each document.

Making a query

The TF-IDF matrix makes it possible to find the documents in the text corpus which has the best correlation with specific terms. Using cosine similarity it can give an overview of which documents are most relevant to look at, in order to make a topic modelling analysis. Here the top ten most relevant documents is shown.

**Topic modelling**

Topic modeling works with a bag-of-words approach, and that a topic can be shown relation between the meaning of each word as a cluster (Ignatow, G. & Mihalcea 2018). It is widely used to discover correlations between discourses in journalism and politics, and in this assignment the goal was to select a corpus relevant to a query, and perform topic modelling of this corpus. As this was not within the competences of the author, the topic modelling is performs on the full corpus. Finding a solution to this problem will be the next step, working with the Guardian API.

**Literature**

Ignatow, G. & Mihalcea, R. (2018) An Introduction to Text Mining : Research Design, Data Collection, and Analysis.Los Angeles, CA. Sage

Greene, D. and Cross, J.P. 2017. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. Political Analysis 25(01), pp. 77–94.

1. http://ufm.dk/publikationer/2014/the-danish-code-of-conduct-for-research-integrity [↑](#footnote-ref-1)
2. <https://dkuni.dk/om-os/internationalt/code-of-conduct/> [↑](#footnote-ref-2)
3. Avisartikel [↑](#footnote-ref-3)
4. <https://ec.europa.eu/programmes/horizon2020/en> [↑](#footnote-ref-4)
5. <https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm> [↑](#footnote-ref-5)
6. <https://www.go-fair.org/fair-principles/> [↑](#footnote-ref-6)
7. <https://www.datatilsynet.dk/generelt-om-databeskyttelse/hvad-er-databeskyttelse/> [↑](#footnote-ref-7)
8. <https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf> [↑](#footnote-ref-8)
9. <https://gist.github.com/fsluis/1fef039e7a4065546daeab9d76dad20b> [↑](#footnote-ref-9)