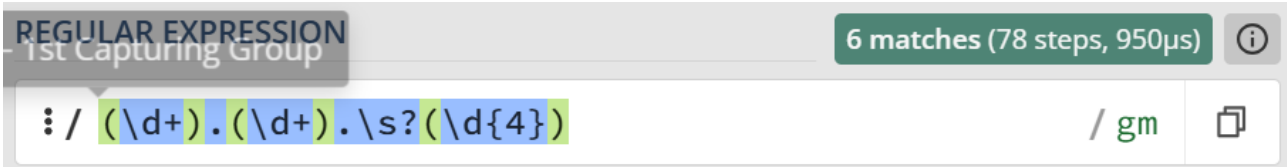
Uge 08 - Regex og OpenRefine

Spørgsmål 1: *What regular expressions do you use to extract all the dates in this blurb: http://bit.ly/regexexercise2 and to put them into the following format YYYY-MM-DD?*

Som det fremgår af spørgsmålet, fremkommer datoerne i datasættet i forskellige formater. Vha. Regular Expressions kan disse datoer formateres, således at de fremstår som YYYY-MM-DD.

Følgende kommando benyttes for at omskrive datoerne:



Ligning 1: screenshot fra https://regex101.com/r/sBGaPT/1

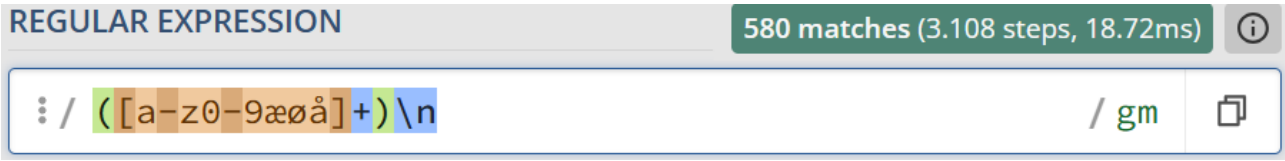
Udtrykket beskriver følgende:

(\d+)	\d: matcher et hvilket som helst tal. +: matcher det førnævnte udtryk (her \d), hvis det fremstår en eller flere gange. Dette fremkommer to gange for at få MM-DD (hhv. gruppe $1 og $2) til at fremkomme. (): gør at hvert udtryk defineres som en gruppe.
.	Matcher alle tegn. Dette gør, at hvad end datoerne i den originale tekst er adskilt af, vil det matche.
\s?	\s: matcher mellemrum eller ny linje; her mellemrum. ?: matcher førnævnte udtryk (her\s), hvis det fremkommer en eller nul gange, da der i nogle tilfælde i den originale tekst indgår mellemrum i datoangivelsen.
(\d{4})	\d: se række 1. {4}: matcher det foregående udtryk (her \d) præcis 4 gange for at få udtryk YYYY (gruppe $3). (): se række 1.

Efterfølgende vælger vi at substituere med udtryk $3-$1-$2 for at datoerne kommer i formatet YYYY-MM-DD. Den endelige rettede tekst kan ses via linket indsat ved: *ligning 1*.

Spørgsmål 2: *Write a regular expression to convert the stopwordlist (list of most frequent Danish words) from Voyant in http://bit.ly/regexexercise3 into a neat stopword list for R (which comprises "words" separated by commas, such as http://bit.ly/regexexercise4).*

Vi har brugt regex for at få den samlede voyant-stopliste (liste adskilt af linjeskift) skrevet om til en R-stopliste (tekst adskilt af "*ord*", og mellemrum). Vi har lavet nedenstående udtryk for at omskrive:



Ligning 2: screenshot fra https://regex101.com/r/x5knWJ/1

([a-z0-9æøå]+)	[]: definerer den liste af tegn, som man ønsker at finde. a-z: matcher ethvert lille bogstav fra a-z. Alle ord med disse bogstaver er inkluderet. 0-9: matcher ethvert tal. æøå: tilføjes manuelt som andre tegn, da det ikke-danske program ikke kender dem som bogstaver i en alfabetisk rækkefølge. +: matcher det førnævnte udtryk (her tal og bogstaver indenfor klammen), hvis det fremstår en eller flere gange. Derfor inkluderer udtrykket alle ord. (): gør, at hvert udtryk defineres som en gruppe.
\n	Matcher en ny linje, da ordene i voyant-listen er adskilt af linjeskift.

Se den rettede liste via linket ved: *ligning 2*.

Vi tilføjer til sidst "$1" i substitutionslinjen, fordi gruppen, som vi har beskrevet ovenfor, alle er ord, der skal omgives af citationstegn, for at den lever op til kravene for en R-stopliste.

Det er nødvendigt at bemærke, at vores regex ikke gør det muligt at inkludere de tilfælde, hvor der indgår andre specialtegn end dem, som vi har inkluderet i ligningen ex. ejefaldet i *cd's*.

Spørgsmål 3: *Then take the stopwordlist from R http://bit.ly/regexexercise4 and convert it into a Voyant list (words on separate line without interpunction)*

Vi bruger stort set samme regex-indtastning som i spørgsmål 2. Der er dog nogle ændringer for at konvertere teksten fra en R-stopliste (tekst adskilt af "*ord*", og mellemrum) til en voyant-stopliste (liste adskilt af linjeskift). Vi vil derfor nedenfor kun beskrive de ændringer, som vi har lavet:

.	Vi har tilføjet tegnet en gang før () og to gange efter (). Det gør, at vi kan fjerne de symboler, som i R-listen står omkring alle ordene.
\n	Har vi fjernet, da R-listen ikke indeholder noget linjeskift, som vi ønsker at slette.

Vi tilføjer til sidst $1\n i substitutionslinjen, fordi gruppen som vi har beskrevet ovenfor, alle er ord, der skal stå på en ny linje, hvis listen skal leve op til kravene for en voyant-stopliste. Se den bearbejdet ordliste her: https://regex101.com/r/5W1mMP/1

Spørgsmål 4: *Does OpenRefine alter the raw data during sorting and filtering?*

OpenRefine ændrer ikke rådata gennem sortering og filtrering uden brugerens involvering. Dog kommer OpenRefine med forslag til sortering og filtrering, der skal godkendes af brugeren.

OpenRefine er et godt værktøj til at sammenslå, sortere og filtrere rådata, hvilket vi tilsammen kan kalde at bearbejde data. Eksempelvis kan vi aktivt vælge at slå kategorier sammen (ex. landsoldat og nationalsoldat), hvis vi vurderer, at det vil understøtte den historiske kontekst eller det spørgsmål, som vi gerne vil svare på. Alt manipulering af rådata kræver tolkning og analyse, og man kan derfor trække dataen i mange retninger og bruge den til mange formål. Filtreringsprocessens formål er altså at forberede datasættet til de næste skridt. Samtidig bliver regnearket langt mere læseligt for iagttageren, når de hurtigt og effektivt skal finde data.

Spørgsmål 5.1: *Fix the interviews dataset in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/dryest by the interviewed farmer households?"*

Datasættet skal bearbejdes, så alle forskellige stavemåder, tegnangivelser osv. ikke forstyrrer aflæsningen af måneder og så hver enkelt måned (og NULL og (blank)) kun forekommer som én kategori.

Ud fra vores redigerede data kan vi konkludere, at de to tørreste måneder i 2016 var september og oktober. De fremkommer hhv. 70 og 74 gange i interviewene som de måneder, hvor der ikke er nok vand. Konklusionen er dog behæftet med stor usikkerhed, da kategorien *months_no_what* i 126 tilfælde er angivet som blank og i 45 tilfælde er angivet som NULL.

Spørgsmål 5.2: *Real-Data Challenge: What are the 10 most frequent occupations "erhverv" among unmarried men or women of 20-30 years in 1801 Aarhus census dataset? (hint: first select either men or women to shrink the dataset to a manageable size, then filter by age, and then use merging to cut the erhvervvariation ruthlessly.)*

De 10 mest forekommende erhverv for ugifte mænd i alderen 20-30 år:

Det er nødvendigt at bemærke, at konklusionen er behæftet med en del usikkerhed. Der er 1880 tilfælde, hvor mandens erhverv ikke er angivet (blank), og samtidig er kategorien *soldat* udtryk for mange forskellige typer af soldat. Mange af mændene er soldat samtidig med, at de varetager et andet erhverv, og det kan være en forklaring på, at der er så mange med netop soldater-erhvervet.

De 10 mest forekommende erhverv for ugifte kvinder i alderen 20-30 år:

Denne konklusion er også behæftet med en del usikkerhed. Størstedelen af kvindernes erhverv falder i kategorien NULL, som vi har lavet som én samlet kategori for alle de angivelser, der ikke er et erhverv (ex. krøbling). Desuden er der 2142 tilfælde, hvor kvindens erhverv ikke er registreret (blank).