

Week8 assignment - Digitale arkiver og metoder

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?*

Jeg bruger regex: `\d{1,2}.\d{1,2}..\d{2,4}` til at finde alle datoerne i de forskellige formater i teksten.

`\d{1,2}` → Starten på mit søgeudtryk. Med det her leder jeg efter enten et eller to tal, der står sammen og uden mellemrum.

`.` → metakarakter der tilgodeser de forskellige formateringer i teksten, såsom '-', '/' eller mellemrum.

`\d{1,2}` → Ligesom den første leder jeg her efter et-to tal, altså måneden eller dagen i en dato.

`..` → Jeg sætter jeg to punktummer, der tillader et tegn mellem dato og år, men også et mellemrum.

`\d{2,4}` → Til sidst leder jeg efter to-fire tal, altså et årstal.

Derefter indrammer jeg de tre `\d{x}` i min regex med parenteser for at gøre dem gruppérbare, så jeg kan substituere dem i teksten i et ensartet format, som jeg bestemmer; nemlig "YYYY-MM-DD".

Min endelige regex ser sådan ud:

`(\d{1,2}).(\d{1,2}).(\d{2,4})`

Og til indsætte dem i formatet "YYY-MM-DD": \$3-\$2-\$1

[Link til min løsning på regex101.com](http://bit.ly/regex101.com)

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

- [Stopord \(Liste til R\)](#)

Til denne øvelse bruger jeg 'søg og erstat' i Word til at erstatte anførelsestegn med ingenting (slette dem) og derefter erstatte kommaer med linjeskift. Jeg får en lang liste af de samme ord, som jeg

gemmer i et nyt dokument, klar til brug som stopordsliste i Voyant:

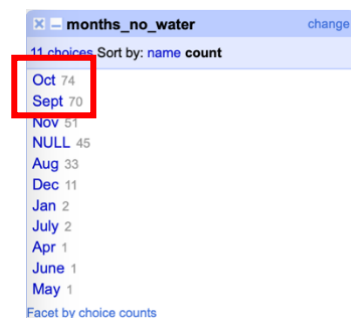
- [Stopord \(Liste til Voyant\)](#)

3. *Does OpenRefine alter the raw data during sorting and filtering?*

Nej. De justeringer man kan lave på data under *filtering* og *sorting* redigerer ikke i rå-dataene. I stedet fungerer de som en slags lag af raffinering af dataene, som man kan positionere, alt efter hvilke informationer man ønsker et overblik over.

4. *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?"*

Månederne oktober og september er de tørreste, hvor hhv. 74 og 70 af de adspurgte farmere nævner dem:



months_no_water	
11 choices. Sort by: name count	
Oct	74
Sept	70
Nov	51
NULL	45
Aug	33
Dec	11
Jan	2
July	2
Apr	1
June	1
May	1

For at nå frem til disse *fixet* data har jeg transformeret indholdet i cellerne i kolonnen "months_no_water" på følgende måde:

0. Create project
1. Text transform on 131 cells in column interview_date: value.toDate()
2. Text transform on 86 cells in column months_no_water: grel:value.replace("[", "")
3. Text transform on 86 cells in column months_no_water: grel:value.replace("]", "")
4. Text transform on 86 cells in column months_no_water: grel:value.replace(" ", "")
5. Text transform on 83 cells in column months_no_water: grel:value.replace(" ", "")

← Her har jeg med fire forskellige GREL-udtryk fjernet de ønskede tegn; [] ' og mellemrum ved at erstatte dem med "ingenting".

5. *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.)*

Ud fra dataene i '1801 Aarhus census' er de 10 mest hyppige erhverv blandt ugifte mænd og kvinder i alderen 20-30 år:

Mænd	Kvinder
1. Soldat (Sammensat gruppe af forskellige former for soldat blandt besvarelserne; nationalsoldat, rekrut, landsoldat, nationalrytter, soldat ved 1. jyske inf. reg., m.v.)	1. Tjenestepige / Husjomfru / Stuepige / Udepige / Husholderske /
2. Bonde (Tjenestekarl, bonde og gårdbeboer, avlskarl)	2. Væverske / Væverpige / Lever af at sy / Spindelkone / Syepige / Spinderske
3. Væver	3. Ikke i job / får almisser / har funktionsnedsættelse af en art: (Almisselem, Vanfør, 'dum og døv', 'vanfør og tåbelig', krøbling, hospitalslem, m.v.)
4. Skrædder	4. Lever af sine midler
5. Matros	5. Går med bagerkurven
6. Skoleholder	6. Gårdbeboer
7. Skriverkarl	N/A
8. Snedker	N/A
9. Skomager	N/A
10. Købmandskarl	N/A

I denne øvelse har jeg filtreret dataene med *text facets* som vist her:



Efterfølgende har jeg brugt funktionen *Cluster and edit* til at (forsøge at) sammenflette de mange variationer af erhverv, der dækker over samme job/jobtype, hvilket har givet mig informationerne til mit svar i ovenstående tabel.