Aflevering 21.2

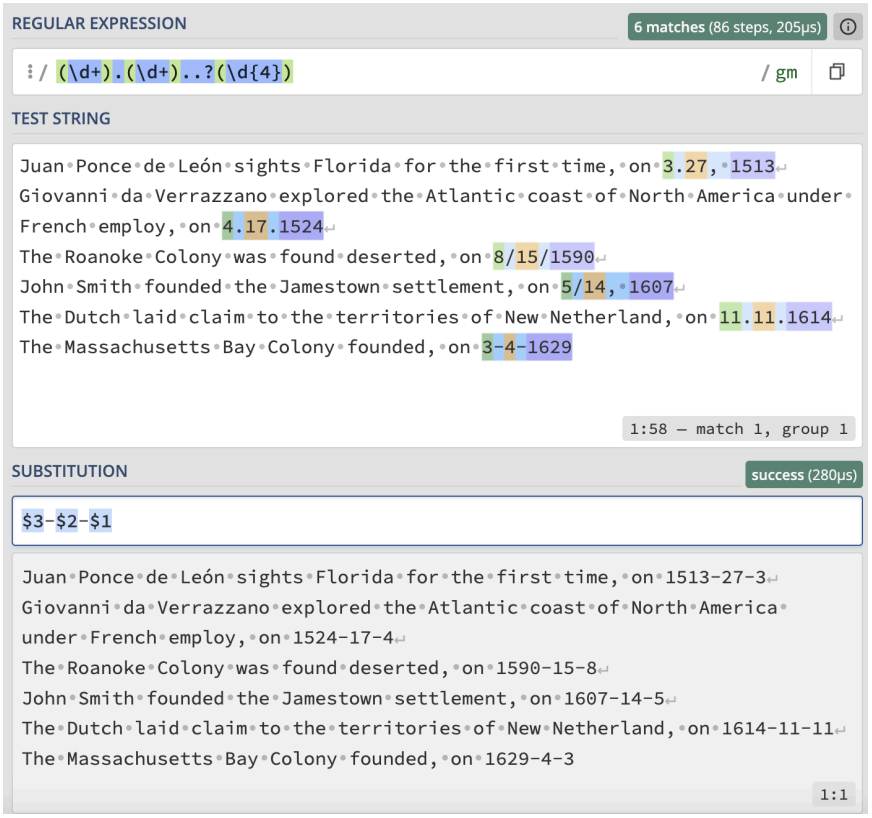
Week 8 assignment is, inevitably, about Regular Expressions and OpenRefine

Upload a text file or a PDF with your answers/solutions to the problems below. Beware of making the submission legible and understandable to another reader; for example, consider using the "Save regex" functionality in regex101.com, which allows you to create a link out of your solution and share the link for easy use by your colleagues. Remember that you can elaborate solutions in groups, but need to submit **individually.**

1. 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 ?

Vi har skrevet en kode for hvert tal vi har fået informeret. Dato, måned og årstal. Med dato og måned har vi brugt "(\d+)." med et punktum efterfølgende for at tage højde for mellemrum. Koden til årstallet bliver "..?(\d+{4})", fordi vi skal tage højde for flere mellemrum mellem måned og årstal, hvilket er et 4-cifret tal, derfor indsat {4}.

Regex101: https://regex101.com/r/sTjRDy/1



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)

Vi har taget og registreret alle vores bogstaver, særbogstaver (danske æ, ø og å). Isolerede det for hver sætning med "\X" og "+" for at matche tidligere elementer. Vi har brugt samme kode før, men gjort det modsatte for at kode indføres og tilføjet ekstra bogstaver, .

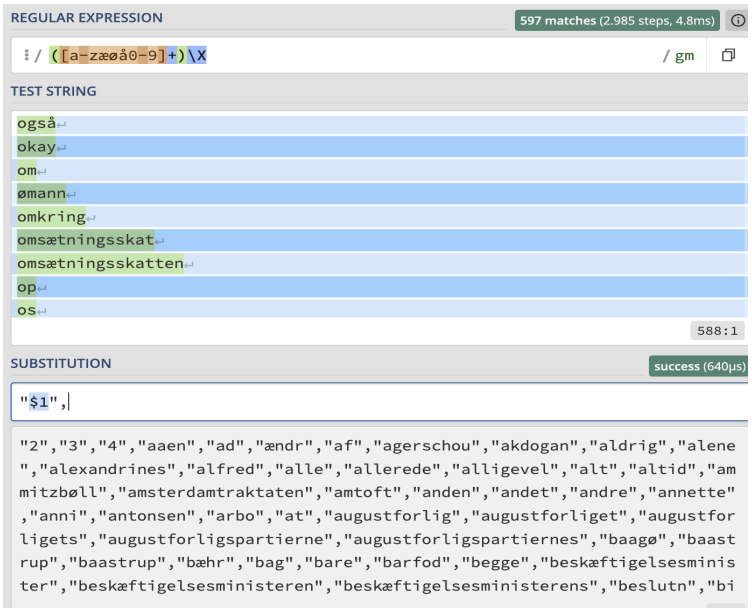
Regex101 (1): Voyant to R:

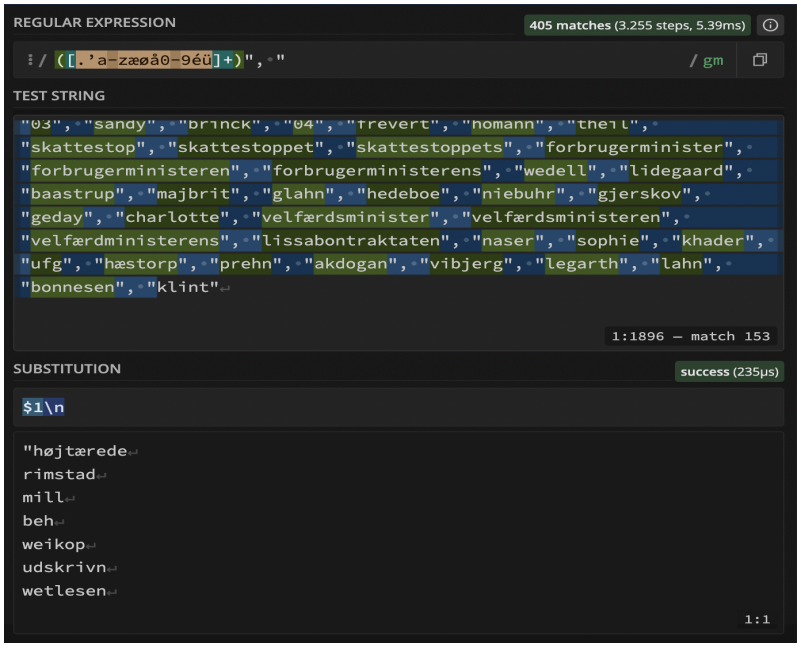
https://regex101.com/r/X894x9/1

Regex101 (2): R to Voyant:

https://regex101.com/r/D9y7x7/1

(1)



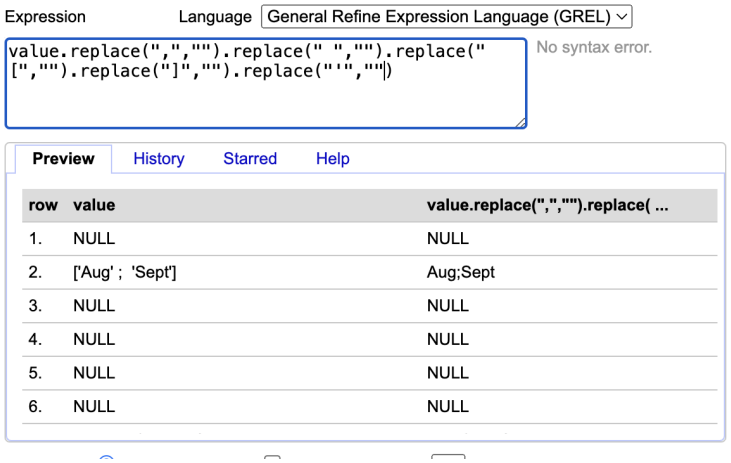
(2) 

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

Nej, det ændrer ikke den rå data, den sætter blot dataen op på en ny måde, hvori man kan manipulere dataen og sætte det op på andre og nye måder, f.eks. at rette stavefejl og finde hyppighed i datasættet. Det er altså præsentationen, der ændrer sig, ikke selve dataen.

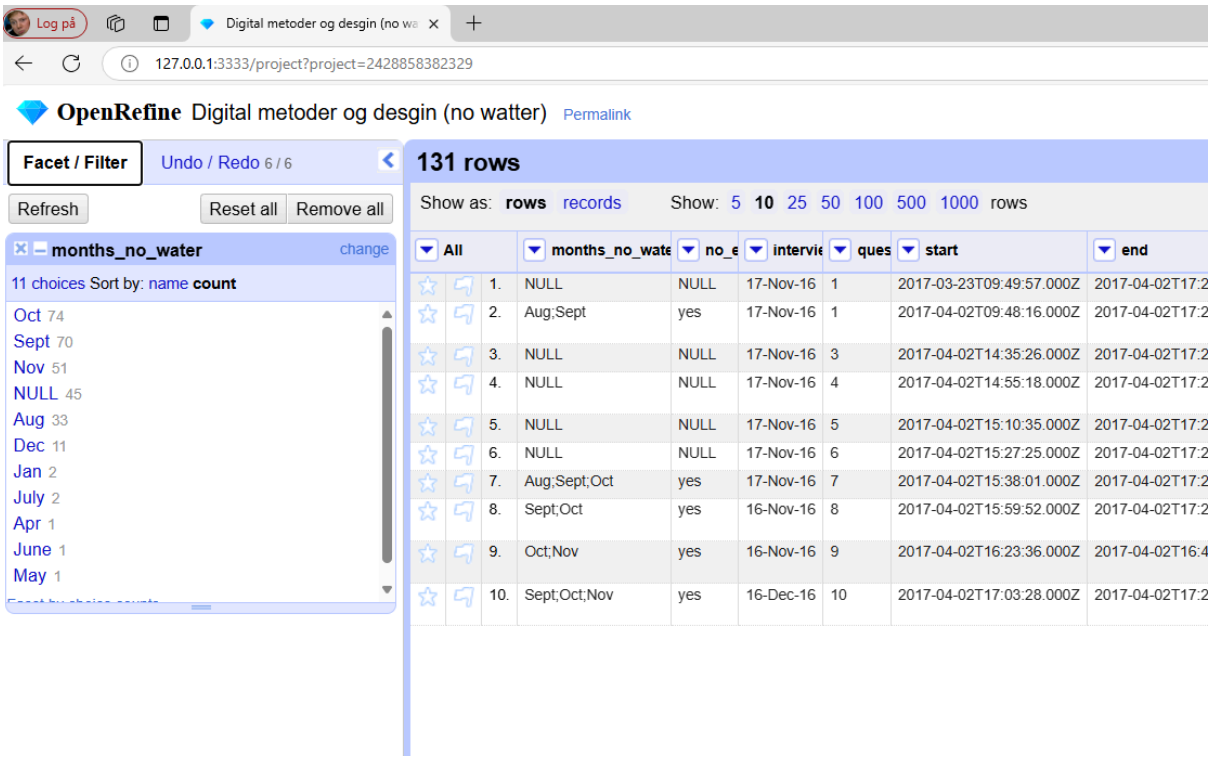
4. Fix the interviews dataset in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/dryest by the interviewed farmer households?"

- vi har lagt tallene ind på openRefine
- så har vi fjernede ‘, [] og mellemrum



- derefter har vi lavet costume facet med value.split(“;”) som dermed splitter de forskellige data

De 2 måneder, hvor der har været tørrest, er okt og september.



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.)

Kvinder:

Har lave følgende ændringer og reguleringer:

- Koen: har valgt kategorien kvinder
- Alder: har ændrede data til numbers, og derefter udvalgt kategori 20-30 årig igennem Facet
- Civilstand: udvalgt følgende Enke og ugift, (blank er ikke inkluderet grundet usikkerhed)

Cluster beslutninger og begrundelse

- Indsidder og inderste, er beskrivelser af en boligsituation og ikke et erhverv, vi betragter det derfor ikke som en kategori. Kategorier som “indsidder og væver” bliver altså blot til væver. Kategorien “indsidder” bliver blot ikke betragtet som et erhverv
- tjener ved forældrene og tjener ved faren osv. er sat sammen til tjener ved forældrene
- vi betragter ikke “husjomfru” som et erhverv

- vi betragter ikke “lever af sine midler”som et erhverv

Listen bliver således:

1. tjenestepige: 31 kvinder
2. væverske: 21 kvinder
3. tjener-forældrene: 12 kvinder
4. spinder: 8 kvinder
5. husholderske: 7 kvinder
6. kokkepige: 5 kvinder
7. bryggerpige: 4 kvinder
8. hospitaslem: 4 kvinder
9. skrædderpige: 4 kvinder
10. Ernærer sig af sygning: 3 kvinder