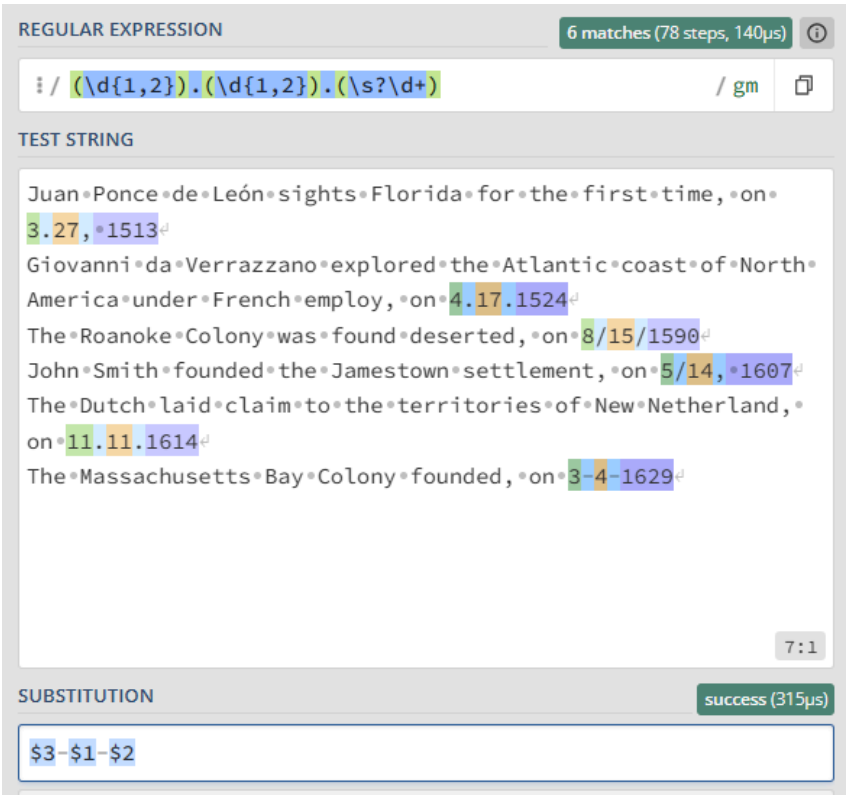
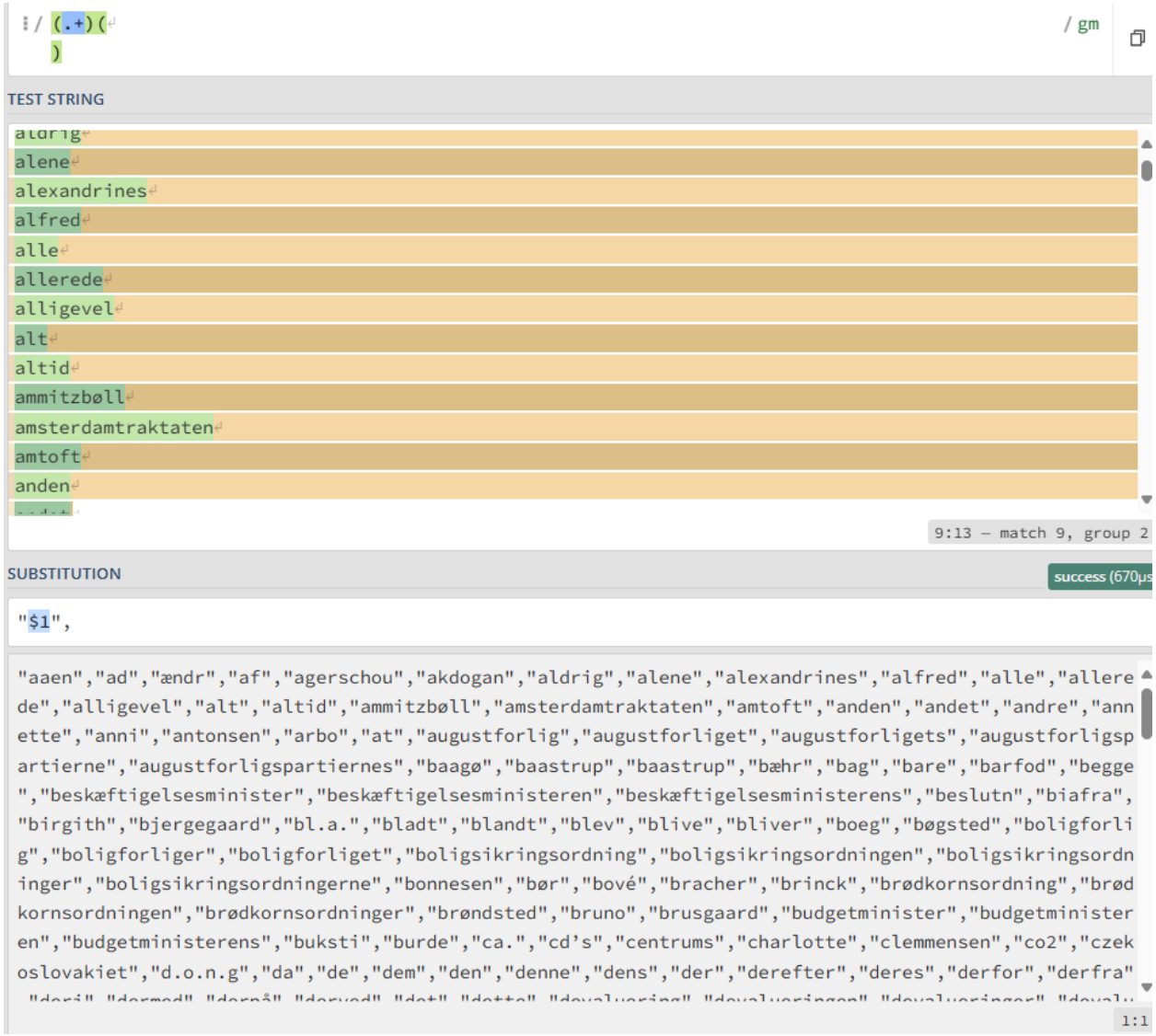
Week 8 assignment is, inevitably, about Regular Expressions and OpenRefine
Upload a text file or a PDF with your answers/solutions to the problems below. Beware of making the submission legible and understandable to another reader; for example, consider using the "Save regex" functionality in regex101.com, 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. 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 ?

regex101: build, test, and debug regex



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)4

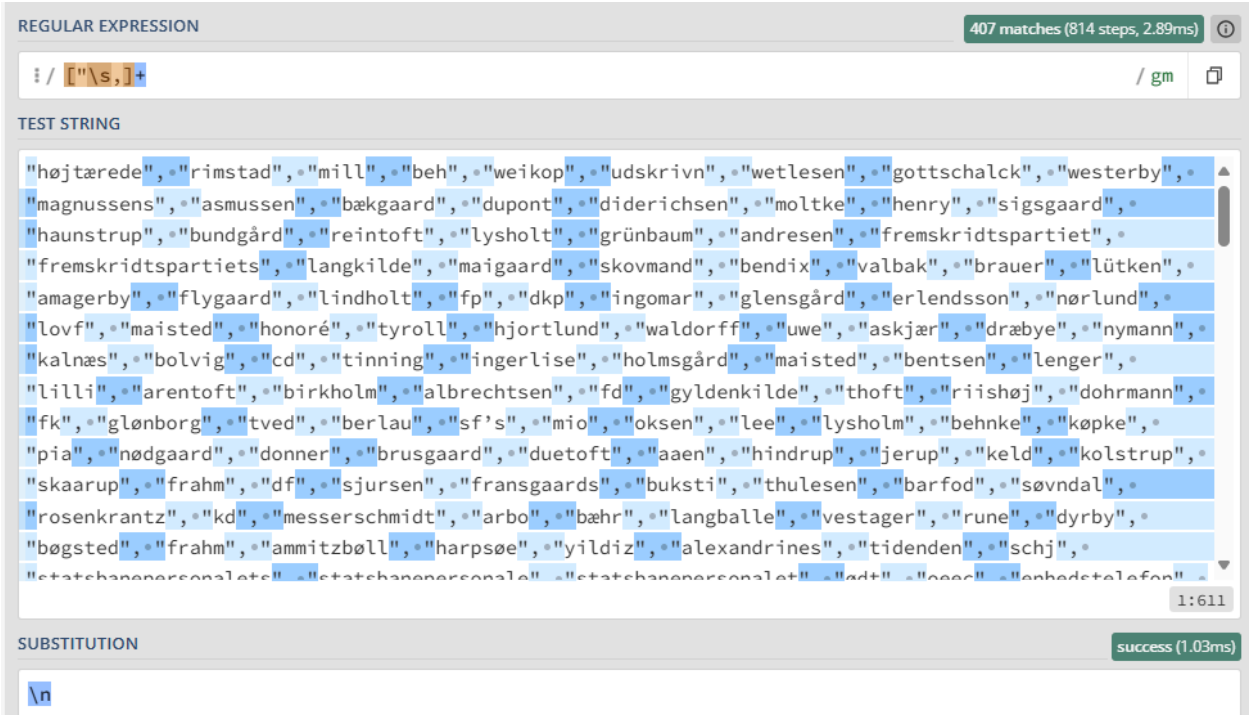


Det vil resultere at alle ordene står som ovenfor, separeret med “værdi”,

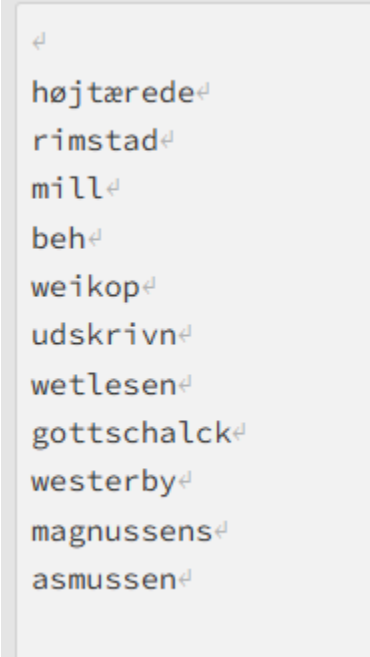
2. Then take the stopwordlist from R http://bit.ly/regexexercise4 and convert it into a Voyant list (words on separate line without interpunction)4

Herefter skal vi tage stop ord listen, og omstrukturere daterne, så vi kan bruge den i voyant. Her kopiere man ordene og sætter dem ind i regex:

Derefter skal man fjerne alle værdierne, som separere ordene:



Denne kodning skulle gerne resultere i at ordene bliver som følgende:



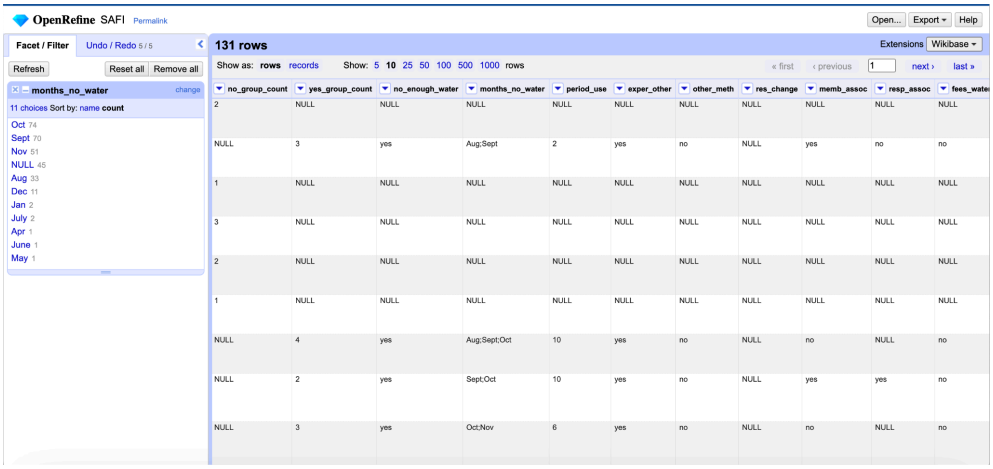
https://regex101.com/r/vGNdvw/1
Herefter skal man kopiere ordene Ctrl-A + Ctrl-C
Så skal man kopiere ordene i en txt fil, og så skulle den gerne være klar til at bruge i voyant

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

Nej, openrefine kan komprimere data til en mere overskuelig sammensætning. Men den laver ikke om på dataene.

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?"

- Først har vi gjort cellerne under “months_no_water” mere overskuelige ved at fjerne “[]” og de mellemrum der var. Det har vi gjort ved at trykke “Edit cell” → “Transform” → Skrive: value.replace(“[“,””).replace(“]”,””).replace(“ ”,””)
- Da flere af cellerne indeholder flere datasvar har vi splittet hvert datasvar og derefter grupperet dem, dog uden at selve dataen i cellerne ændrer sig. Det har vi gjort ved at trykke “Facet” → “Costum text facet” → Skrive: value.split(“;”)
- Vi kan herefter så aflæse hvilke to måneder i datasættet der var tørrest ved at aflæse hvilke to datasvar der er rapporteret flest gange
- Ud fra dette kan man altså se at det er oktober og september der har været de to tørreste måneder



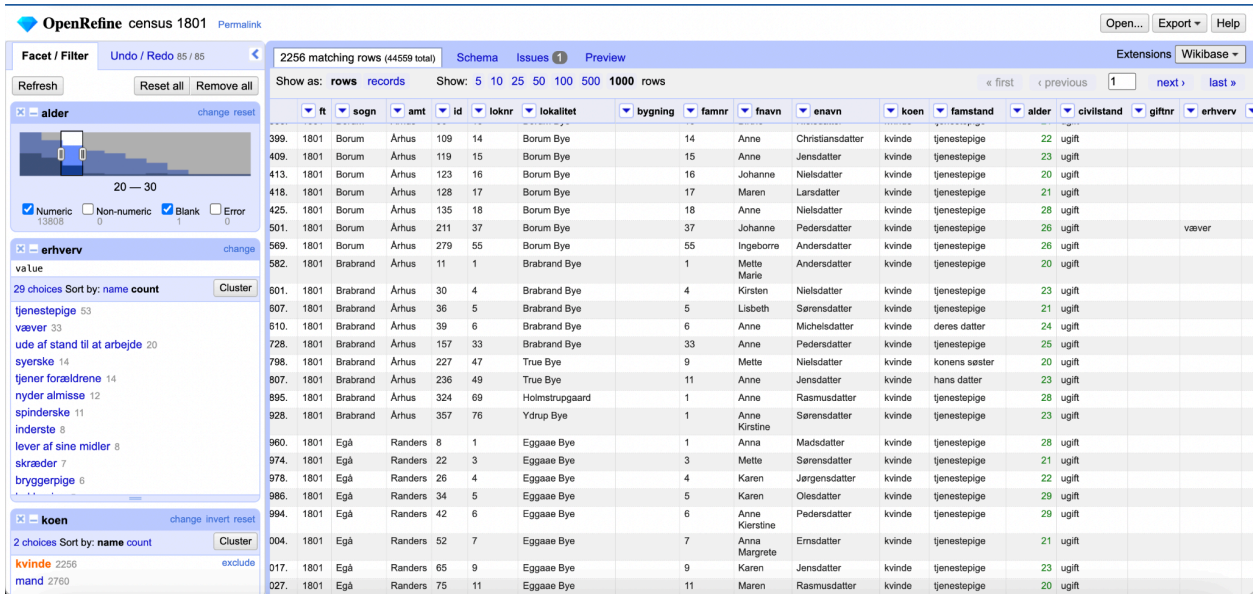
5. Real-Data Challenge: What are the 10 most frequent occupations "erhverv" among unmarried men or women of 20-30 years in 1801 Aarhus census dataset? (hint: first select either men or women to shrink the dataset to a manageable size, then filter by age, and then use merging to cut the erhvervvariation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

- Vi har valgt at kigge på ugifte kvinder i datasættet og har derfor kun inkluderet de data der er hvor “koen” er markeret som “kvinde” og “civilstand” markeret med “ugift” eller “enke” . Da opgaverne er stillet med fokus på alderen 20-30 år, er der altså kun inkluderet data, hvor både “koen” er markeret som “kvinde” og “alder” er markeret med data i intervallet [20;30]. Herefter har vi grupperet dataene, så vi har så få grupperinger, at det er muligt at give et mere overskueligt svar. Grupperingen af dataen er gjort ved

key collision, methaphone 3, og nearest neighbor, levensthein. Her har vi vurderet, hvilken data der var relevant at gruppere. Herefter har vi manuelt grupperet den data, der ikke var mulig at gruppere ved hjælp af Openrefines funktioner, hvilket vi har gjort ved at ændre navnet på den data vi har ville tilføje til en gruppe, nogen af dem havde flere "erhverv" så her valgte vi selektivt hvor de passet bedst ind. Vi har valgt at besvare spørgsmålet med 29 forskellige datasvar, men det ville helt klart være muligt at gruppere dataen yderligere. Dette ville dog indebære en grovere sortering af dataen og dermed også et mere upræcist svar.

- På baggrund af ovenstående kan vi derfor konkludere, at de 10 hyppigste erhverv blandt ugifte kvinder i alderen 20-30 år var:
    1) tjenestepige
    2) væver
    3) ude af stand til at arbejde
    4) syerske
    5) tjener forældrene
    6) nyder almisse
    7) spinderske
    8) inderste
    9) lever af sine midler
    10) skræder

Svarene fra Openrefine kan ses i screenshot nedenunder



-

```
tjenestepige    53
væver   33
ude af stand til at arbejde     20
syerske 14
tjener forældrene       14
nyder almisse   12
spinderske      11
inderste        8
lever af sine midler    8
skræder 7
bryggerpige     6
kokkepige       5
arbejdsløs      2
gaardbeboer     2
har alt frit paa gaarden        2
huusbeboer med jord     2
afdød selveier anders jensens huus med jord     1
amme    1
arbejder med håndarbejde        1
barnepige       1
daglejer        1
giør stoele     1
går med bagerkurven     1
huusmoder       1
i besøg på gården       1
inderste der ellers er udi fast tieneste        1
```

```
lever af husnæring      1
mejerske        1
opholdende      1
(blank) 2044
```