

# 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. 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>

The screenshot shows the regex101.com interface. At the top, it says "REGULAR EXPRESSION" and "6 matches (86 steps, 205µs)". The regular expression entered is `/(\d+).(\d+).?(\d{4})/gm`. Below this, the "TEST STRING" section contains a text blurb about the discovery of Florida. The dates in the text are highlighted with colored boxes: "3.27.1513", "4.17.1524", "8/15/1590", "5/14.1607", "11.11.1614", and "3-4-1629". At the bottom, the "SUBSTITUTION" section shows the result of the regex: `$3-$2-$1`. The text blurb is shown again with the dates formatted as YYYY-MM-DD: "1513-27-3", "1524-17-4", "1590-15-8", "1607-14-5", "1614-11-11", and "1629-4-3".

REGULAR EXPRESSION 6 matches (86 steps, 205µs)

`/(\d+).(\d+).?(\d{4})/gm`

TEST STRING

Juan Ponce de León sights Florida for the first time, on 3.27.1513  
Giovanni da Verrazzano explored the Atlantic coast of North America under French employ, on 4.17.1524  
The Roanoke Colony was found deserted, on 8/15/1590  
John Smith founded the Jamestown settlement, on 5/14.1607  
The Dutch laid claim to the territories of New Netherland, on 11.11.1614  
The Massachusetts Bay Colony founded, on 3-4-1629

1:58 — match 1, group 1

SUBSTITUTION success (280µs)

`$3-$2-$1`

Juan Ponce de León sights Florida for the first time, on 1513-27-3  
Giovanni da Verrazzano explored the Atlantic coast of North America under French employ, on 1524-17-4  
The Roanoke Colony was found deserted, on 1590-15-8  
John Smith founded the Jamestown settlement, on 1607-14-5  
The Dutch laid claim to the territories of New Netherland, on 1614-11-11  
The Massachusetts Bay Colony founded, on 1629-4-3

1:1

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, .

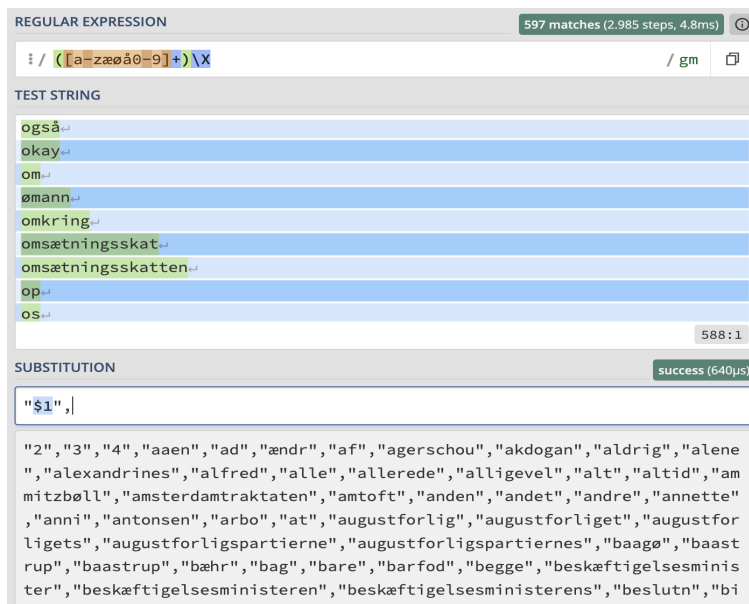
Regex101 (1): Voyant to R:

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

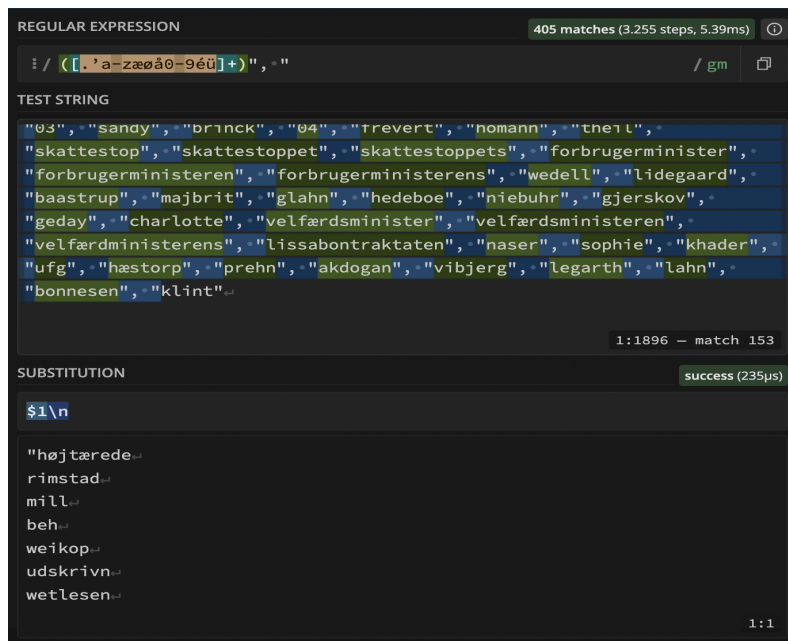
Regex101 (2): R to Voyant:

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

(1)



(2)

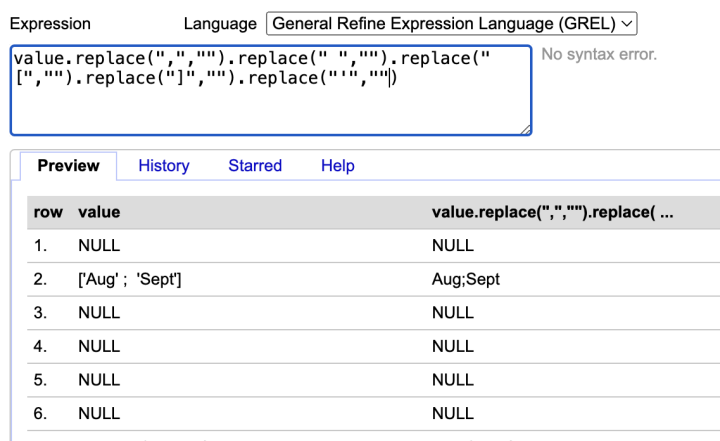


### 3. Does OpenRefine alter the raw data during sorting and filtering?

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 stavfejl 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.

The screenshot shows the OpenRefine web interface. On the left, a facet titled 'months\_no\_water' is displayed with 11 choices, sorted by name and count. The choices are: Oct (74), Sept (70), Nov (51), NULL (45), Aug (33), Dec (11), Jan (2), July (2), Apr (1), June (1), and May (1). The main table displays 131 rows of data. The columns are: All, months\_no\_wate, no\_e, intervie, ques, start, and end. The table shows 10 rows of data, including NULL values and specific dates.

|     | All          | months_no_wate | no_e      | intervie | ques                     | start           | end |
|-----|--------------|----------------|-----------|----------|--------------------------|-----------------|-----|
| 1.  | NULL         | NULL           | 17-Nov-16 | 1        | 2017-03-23T09:49:57.000Z | 2017-04-02T17:2 |     |
| 2.  | Aug;Sept     | yes            | 17-Nov-16 | 1        | 2017-04-02T09:48:16.000Z | 2017-04-02T17:2 |     |
| 3.  | NULL         | NULL           | 17-Nov-16 | 3        | 2017-04-02T14:35:26.000Z | 2017-04-02T17:2 |     |
| 4.  | NULL         | NULL           | 17-Nov-16 | 4        | 2017-04-02T14:55:18.000Z | 2017-04-02T17:2 |     |
| 5.  | NULL         | NULL           | 17-Nov-16 | 5        | 2017-04-02T15:10:35.000Z | 2017-04-02T17:2 |     |
| 6.  | NULL         | NULL           | 17-Nov-16 | 6        | 2017-04-02T15:27:25.000Z | 2017-04-02T17:2 |     |
| 7.  | Aug;Sept;Oct | yes            | 17-Nov-16 | 7        | 2017-04-02T15:38:01.000Z | 2017-04-02T17:2 |     |
| 8.  | Sept;Oct     | yes            | 16-Nov-16 | 8        | 2017-04-02T15:59:52.000Z | 2017-04-02T17:2 |     |
| 9.  | Oct;Nov      | yes            | 16-Nov-16 | 9        | 2017-04-02T16:23:36.000Z | 2017-04-02T16:4 |     |
| 10. | Sept;Oct;Nov | yes            | 16-Dec-16 | 10       | 2017-04-02T17:03:28.000Z | 2017-04-02T17:2 |     |

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 år igennem Facet
- Civilstand: udvalgt følgende Enke og ugift, (blank er ikke inkluderet grundet usikkerhed)

### Cluster beslutninger og begrundelse

- Indsiddet og inderste, er beskrivelser af en boligsituation og ikke et erhverv, vi betragter det derfor ikke som en kategori. Kategorier som "indsiddet og væver" bliver altså blot til væver. Kategorien "indsiddet" 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

OpenRefine UNI opgave 1 [Permalink](#)

Facet / Filter Undo / Redo 5 / 5

2,256 matching rows (44,559 total)

Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first

Refresh Reset all Remove all

**alder** change reset

20 — 30

☒ Numeric 13808 ☒ Non-numeric 0 ☒ Blank 1 ☐ Error 0

**civilstand** change invert reset

3 choices Sort by: name count Cluster

enke 12 exclude

gift 1294 exclude

ugift 2244 exclude

(blank) 20

Facet by choice counts

**erhverv** change

85 choices Sort by: name count Cluster

tjenestepige 31

væverske 21

tjener-forældrene 12

Hus-jomfrue 10

inderste 8

spinder 8

husholderske 7

lever af sine midler 6

kokkepige 5

Bryggerpige 4

hospitalslem 4

skrædderpige 4

Ernærer sig af sygning 3

|     | All  | ft   | sogr  | amt | id | lokn     | lokalit | bygr | famr           | fnavn          | enavn  | koer          | famstand | alder | civil | gift |
|-----|------|------|-------|-----|----|----------|---------|------|----------------|----------------|--------|---------------|----------|-------|-------|------|
| 10. | 1801 | Alrø | Århus | 10  | 1  | Alrø Bye |         | 1    | Eise           | Jeppesdatter   | kvinde | tjenestepige  | 27       | ugift |       |      |
| 11. | 1801 | Alrø | Århus | 11  | 1  | Alrø Bye |         | 1    | Inger          | Jespersdatter  | kvinde | tjenestepige  | 23       | ugift |       |      |
| 12. | 1801 | Alrø | Århus | 12  | 1  | Alrø Bye |         | 1    | Dorthe Sophie  | Nielsdatter    | kvinde | tjenestepige  | 22       | ugift |       |      |
| 13. | 1801 | Alrø | Århus | 13  | 1  | Alrø Bye |         | 1    | Inger          | Jørgensdatter  | kvinde | tjenestepige  | 20       | ugift |       |      |
| 24. | 1801 | Alrø | Århus | 24  | 2  | Alrø Bye |         | 2    | Anne Kierstine | Simonsdatter   | kvinde | tjenestepige  | 28       | ugift |       |      |
| 25. | 1801 | Alrø | Århus | 25  | 2  | Alrø Bye |         | 2    | Eise           | Jespersdatter  | kvinde | tjenestepige  | 25       | ugift |       |      |
| 35. | 1801 | Alrø | Århus | 35  | 3  | Alrø Bye |         | 3    | Maren          | Sørens datter  | kvinde | Tieniste Pige | 23       | ugift |       |      |
| 49. | 1801 | Alrø | Århus | 49  | 5  | Alrø Bye |         | 5    | Maren          | Jens datter    | kvinde | deres datter  | 24       | ugift |       |      |
| 72. | 1801 | Alrø | Århus | 72  | 8  | Alrø Bye |         | 8    | Justine        | Wilhelmsdatter | kvinde | hendes datter | 22       | ugift |       |      |
| 81. | 1801 | Alrø | Århus | 81  | 9  | Alrø Bye |         | 9    | Anne           | Gjerdtsdatter  | kvinde | Tieniste Pige | 24       | ugift |       |      |