

Aflevering 21.2

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

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>

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 stopwordslist (list of most frequent Danish words) from Voyant in <http://bit.ly/regexexercise3> into a neat stopwords list for R (which comprises "words" separated by commas, such as <http://bit.ly/regexexercise4>). Then take the stopwordslist 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, .

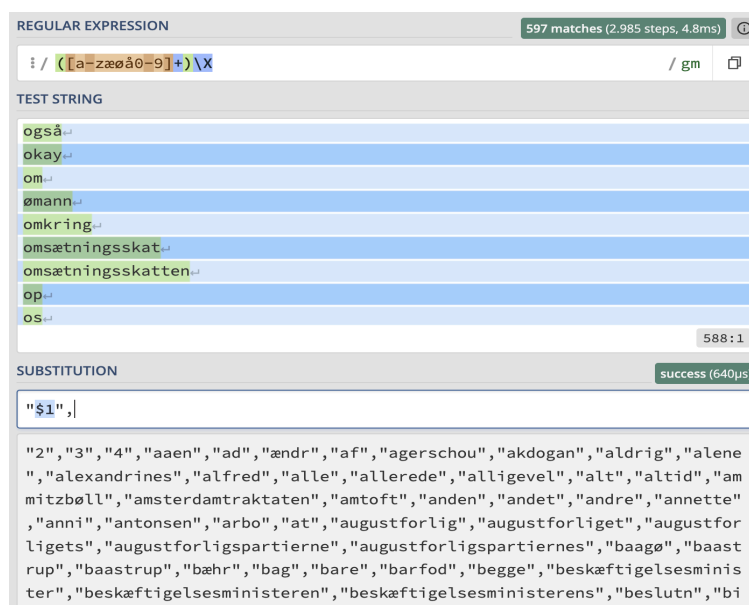
Regex101 (1): Voyant to R:

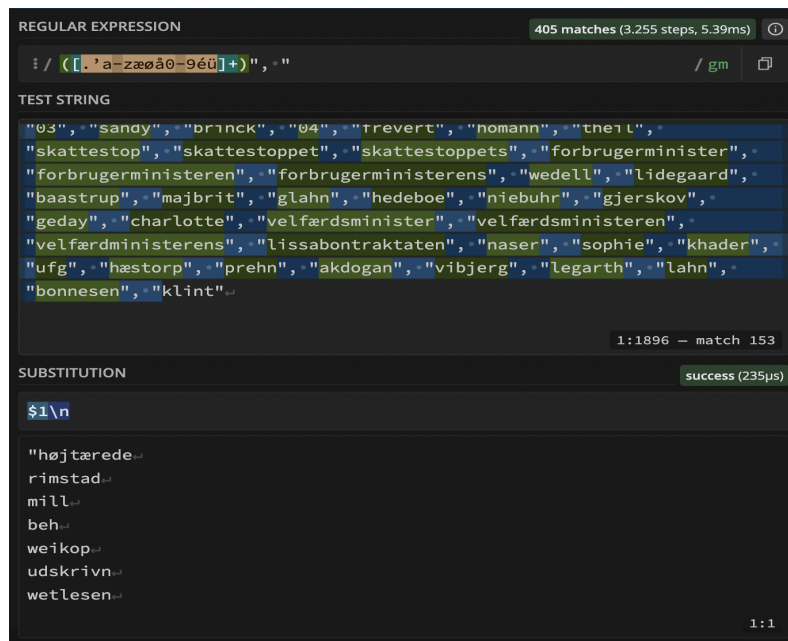
<https://regex101.com/r/X894x9/1>

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 fjernet ' , [] og mellemrum

Expression Language General Refine Expression Language (GREL) ▾

value.replace(",","").replace(" ","").replace("[", "").replace("]", "").replace("'", "")

No syntax error.

Preview History Starred Help

row	value	value.replace(",","").replace(" ...
1.	NULL	NULL
2.	['Aug'; 'Sept']	Aug;Sept
3.	NULL	NULL
4.	NULL	NULL
5.	NULL	NULL
6.	NULL	NULL

- derefter har vi lavet costume facet med `value.split(",")` som dermed splitter de forskellige data.

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

The screenshot shows the OpenRefine web interface. On the left, a facet for 'months_no_water' is displayed with 11 choices, sorted by name 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 shows 131 rows. The first 10 rows are visible, showing columns for row number, months_no_water, no_e, interview, ques, start, and end. The data in the table is as follows:

	months_no_water	no_e	interview	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 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

Facet / Filter

Undo / Redo 5 / 5

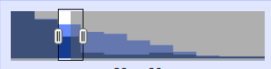
Refresh

Reset all

Remove all

alder

change reset



☒ Numeric 13808

☒ Non-numeric 0

☒ Blank 1

☐ Error 0

civilstand

change invert reset

3 choices Sort by: name count

Cluster

enke 12

gift 1294

ugift 2244

(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

hospitalslæge 4

ekspedient 4

2,256 matching rows (44,559 total)

Show as: rows records

Show: 5 10 25 50 100 500 1000 rows

« first