

# DAM Assignment - week 8

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

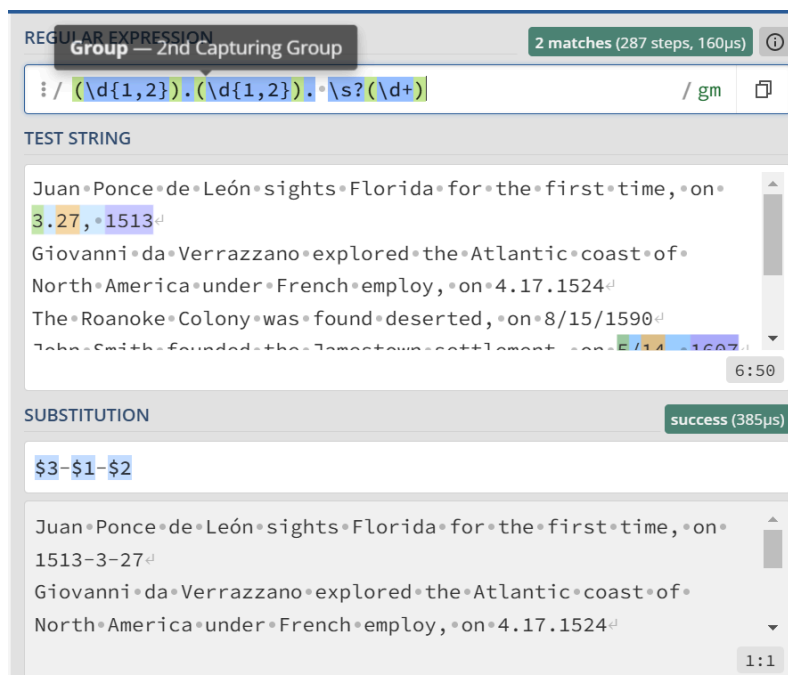
Upload a text file or a PDF with your answers/solutions to the problems below. Beware of making the submission legible and understandable to another reader; for example, consider using the "Save regex" functionality in regex101.com, which allows you to create a link out of your solution and share the 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 ?**

**Regular expression:** `(\d{1,2}).(\d{1,2}). \s?(\d+)`

**Substitution:** `$3-$1-$2`



**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 stopwords list for R (which comprises "words" separated by commas, such as <http://bit.ly/regexexercise4> ). Then**

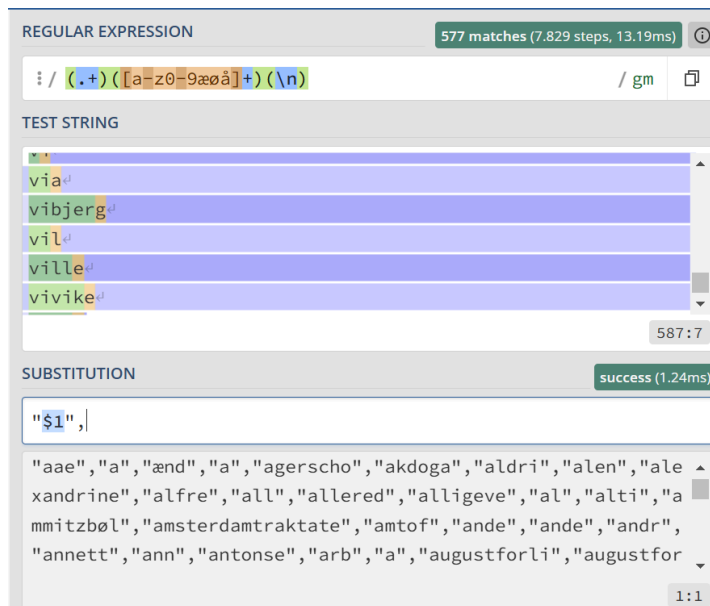
Lasse, Sofie og Emilie

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

### **Voyant stopwordslist -> R stopwordslist**

**Regular expression:** `(.+)([a-z0-9æøå]+)(\n)`

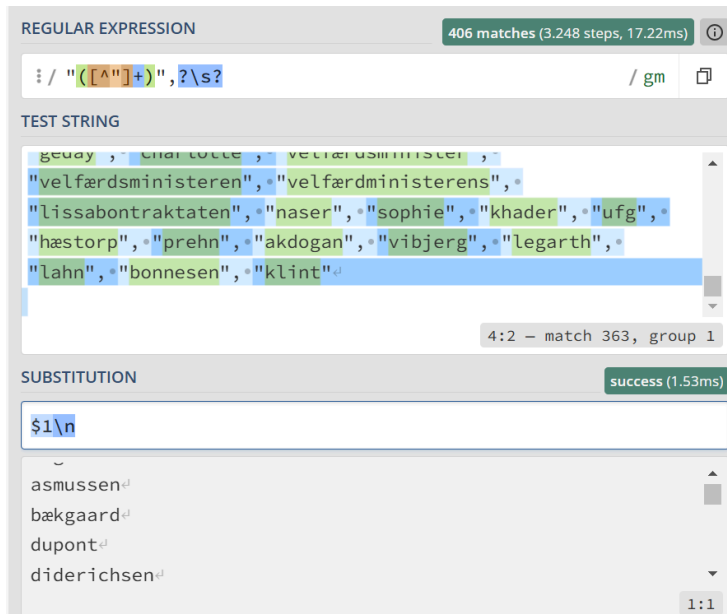
**Substitution:** `"$1",`



### **R stopwordslist -> Voyant stopwordslist**

**Regular expression:** `"([^\"]+)",?\s?`

**Substitution:** `$1\n`



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

**Answer:** No, OpenRefine only cleans up the data aligned with our regular expressions, so that we can find the values that we are looking for. The raw data remains the same throughout the process.

### 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?"



```
jan: 2
apr: 1
may: 1
june: 1
july: 2
aug: 1+4+10+11+6+1 = 33
sep: 1+4+10+11+6+1+21+14+1 = 69
oct: 1+4+10+11+1+1+8+1+8+2+1+21+14+1 = 84
nov: 1+4+10+1+1+1+8+1+1+21+1 = 49
dec: 1+4+2+1+2+1+1 = 12
```

From the data above we can constitute that October and September are the driest months

**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.)**

10 most frequent occupations for **unmarried men of 20-30 years** :

Name of occupation	Frequencies
Tjenestekarl	1.591
Soldat	513
Væver	40
Skræder	24
Matros	20
Læredreng/elev	20
Lærer	20
Gårdmand/bonde	15
Skræddersvend	14
Snedker	11

10 most frequent occupations for **unmarried women of 20-30 years**:

Name of occupation	Frequencies
Tjenestepige	1.495
Væverske	36
Husjomfrue	21
Spinder	17
Husholderske	11
Syrerske	10
Lever af sine midler	8

Lasse, Sofie og Emilie

Mejeripige	8
Skræderpige	7
Kokkepige	6

**Almisse: 20.** We have chosen not to include this under ‘erhverv’ because it is not an actual job despite the recipient receiving “wages”. Nonetheless, we have decided to highlight it anyway because it could be useful for a possible later comparison between this dataset and another from a later date.