**Week 8 assignment 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*](http://bit.ly/regexexercise2) *and to put them into the following format YYYY-MM-DD ?*

We used this regular expression: (\d{1,2}).(\d{1,2}).\s?(\d{2,4})

Here is a link to our solution: <https://regex101.com/r/pACQCa/1>

1. *Write a regular expression to convert the stopwordlist (list of most frequent Danish words) from Voyant in* [*http://bit.ly/regexexercise3*](http://bit.ly/regexexercise3) *into a neat stopword list for R (which comprises "words" separated by commas, such as* [*http://bit.ly/regexexercise4*](http://bit.ly/regexexercise4) *). Then take the stopwordlist from R* [*http://bit.ly/regexexercise4*](http://bit.ly/regexexercise4) *and convert it into a Voyant list (words on separate line without interpunction)*

We used this regular expression to convert the stopwordlist from Voyant into a neat stopwordlist for R: (.+)(\n)

Link to our solution: <https://regex101.com/r/DB3cCW/1>

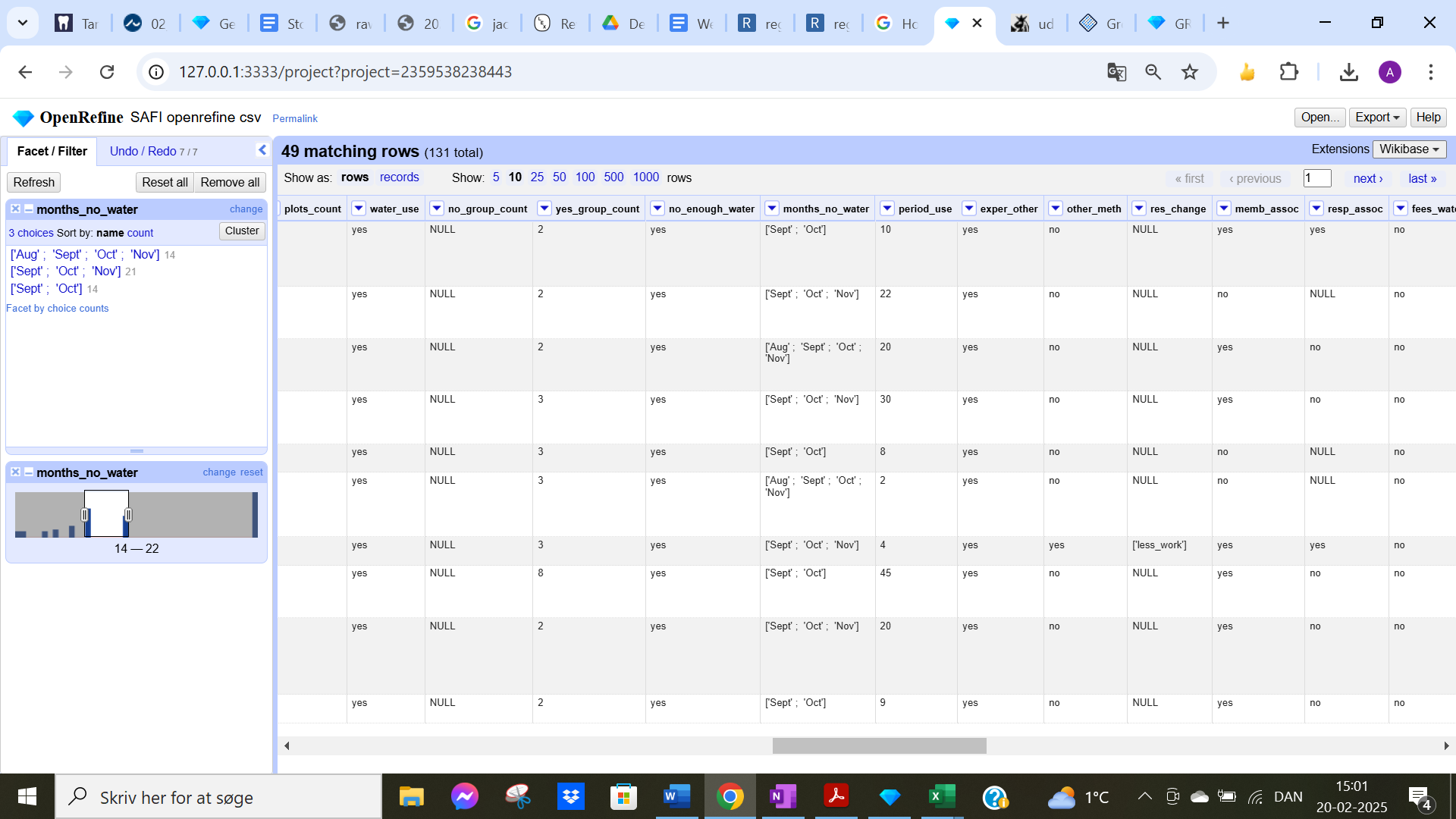
We used this regular expression to convert the stopwordlist from R and into a Voyant list: ([a-z0-9æøå"éè'ü.]+)(. )

Link to our solution: <https://regex101.com/r/MIYM4J/1>

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

Openrefine does not alter the raw data. As a user, you can manipulate the data and sort it into better defined categories. But the raw data will always be available, as you use the “undo/redo” function, and can go back as many steps as you’d like.

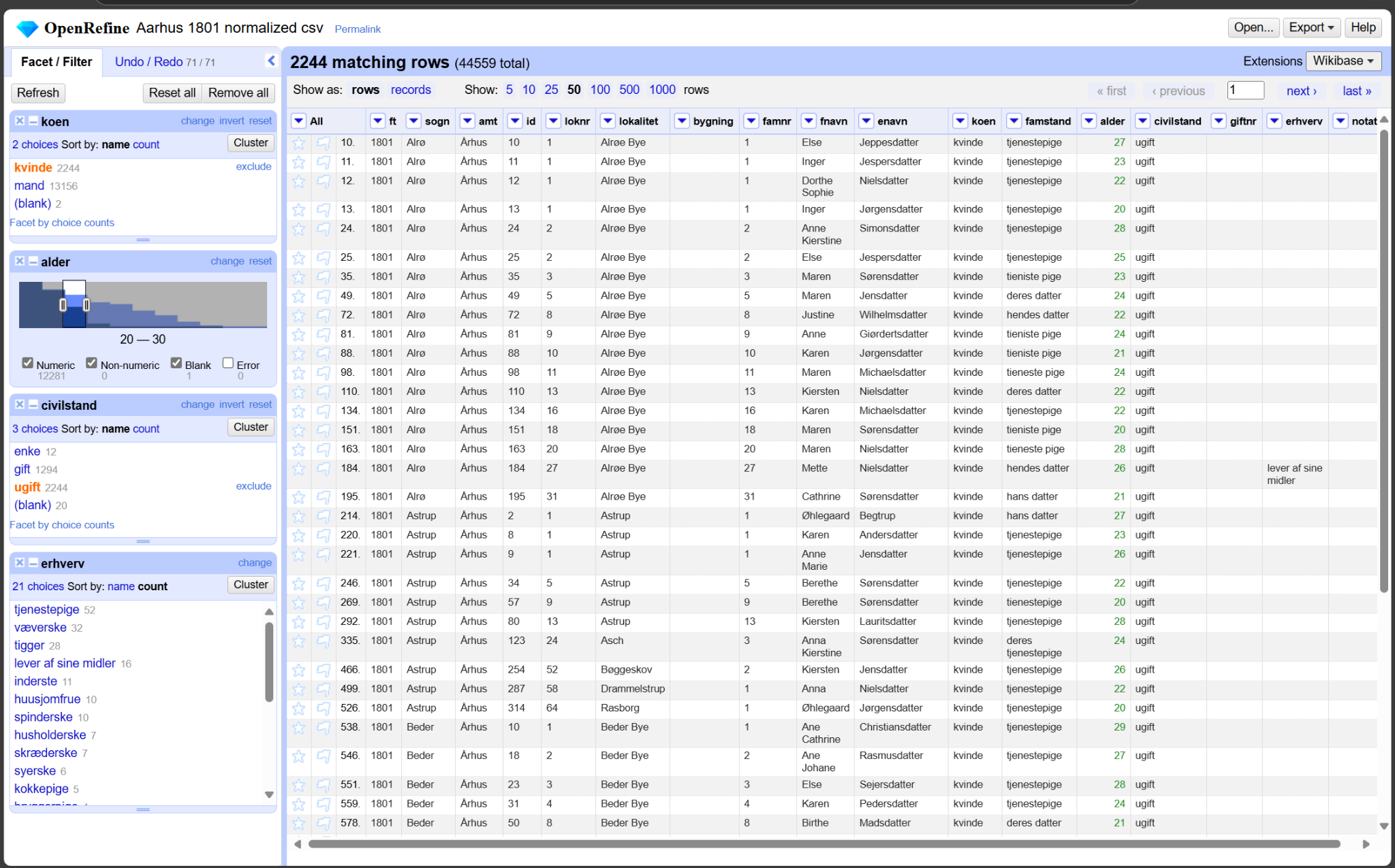
* 4. *Fix the* [*interviews dataset*](https://ndownloader.figshare.com/files/11502815) *in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/dryest by the interviewed farmer households?"*

**

As seen on the screenshot, the two months that are reported most often under the category “months\_no\_water” are September and October. Therefore, the two dryest months are September and October.

* *5. Real-Data Challenge: What are the 10 most frequent occupations "erhverv" among unmarried men or women of 20-30 years in* [*1801 Aarhus*](https://raw.githubusercontent.com/aarhusstadsarkiv/datasets/master/censuses/1801/census-1801-normalized.csv) *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.)*

We selected women and filtered the age to 20-30 years old. Then we merged the professions:



We can now see that the 10 most occurring professions we found regarding women between ages 20-30 are: tjenestepige, væverske, tigger, lever af sine egne midler, inderste, husjomfru, spinderske, husholderske, skrædderske, syerske, kokkepige. But by far the most occurring one is “blank”.