Week 2 - OpenRefine

Emil Bæk Mogensen

2023-09-14

Packages

"'{r} library(tidyverse) library(tinytex) library(knitr) ""

1) **Create a *tidy* spreadsheet/table listing the names of Danish monarchs with their birth- and death-date and duration of reign. They should be sortable by year of birth. Suitable source websites are here and here, but you can also use another source, provided you reference it.**

The source I decided to use was the one from the official website of the Royal Danish House, which has a far deeper description of the monarchs compared to the English url. As it was not possible to upload the data directly into OpenRefine with the URL option, I decided to webscrape the data. Afterwards I cleaned/tidy'ed it to a point where i could finish it manually in excel:. The script to track these steps up until the manual finishing can be found in my GitHub Lastly, i decided to remove interregnums and periods where the throne was contested by multiple kings in the spreadsheet.

2) **Does OpenRefine alter the raw data during sorting and filtering?**

Sorting and filtering does not alter the data. OpenRefine maintains the integrity of the raw data as it simply allows transformations and editing, which in itself allows experimentation without fear of getting lost in the approach or ruining the data.

3) **Fix the interviews dataset in Openrefine enough to answer this question: Which two months are reported as the most water-deprived/driest by the interviewed farmer households?**

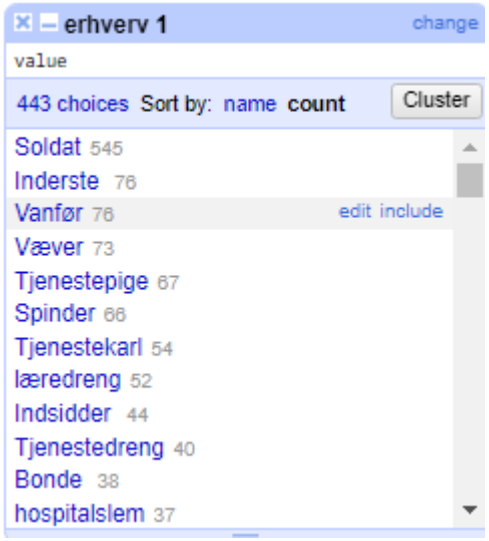
After following and doing the carpentry lesson for the social sciences' i found that the most water-deprived months reported by the farmer households were October and September with 74 and 70 reports respectively. Firstly, I created a custom text facet using the value.replace function to clean the dataset of "[[,." leaving only the abbreviation "xxx" separated by a semicolon of the months in the column. Secondly, I used the "split multi-varied cells" with semicolon separator. Lastly, I clustered the words and got the following picture.
The step-by-step approach can be found at my GitHub.

4) **OPTIONAL Real-Data-Challenge: What are the 10 most frequent occupations (erhverv) among unmarried men and women in 1801 Aarhus? (hint: some expert judgement interpretation is necessary)**

I began by using the URL upload option in Openrefine. Firstly, I applied a text filter with the word "ugift" on the "civilstand" column. Secondly, I decided to split the column with the regex:

```
# ,|\bog
```

This gave me three columns total, of which i focused on the "erhverv 1" column because it is indicated, and thus highly assumed, that this is peoples primary occupancy. Thirdly, I created text facet for the "Erhverv column." I decided to cluster by occupancy in the broadest sense e.g. that of "soldier" for anyone within the military, thus leaving out company, rank and specialty within occupancy. After not being able to do any more of the clustering options, i got the following result.



Although I was asked to focus on the "erhverv" column, I realized by going through it that there were huge discrepancies in the way people responded. This lead me to looking through neighbouring columns, as I was reminded by the carpentry lessons that these mistakes were possible. I found that by clustering the

"famstand" column, although most rows were related to the actual column headline, it had many rows filled with occupancy. Although i swam into deep water and ran into multiple errors and unresponsiveness from OpenRefine in my quest to uncover it, there is no doubt that the occupancy of servant for both men and females respectively was the most frequent. The step-by-step approach to the imported data set can be found at my GitHub.