HW-2 W35 Open Refine

Author: Mina Almasi
Date: 1 of September, 2022

(1) Create a spreadsheet listing the names of Danish monarchs with their birth- and death-date and start and end year of reign. Make it *tidy*! 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. (Group collaboration is expected and welcome. Remember to attach this spreadsheet to Brightspace submission)

The spreadsheet was made in collaboration with Niels Krogsgaard, Jørgen Højlund Wibe & Otto Sejrskild Santesson. Some small personal edits were made.

The spreadsheet is pictured below (also attached).

danish_monarch	birth_year	birth_date	death_year	death_date	start_reign	end_reign
Frederik 5.	1723	31/03/1723	1766	14/01/1766	1746	1766
Christian 7.	1749	29/01/1749	1808	13/03/1808	1766	1808
Frederik 6.	1768	28/1/1768	1839	3/12/1839	1808	1839
Christian 8.	1786	18/09/1786	1848	20/01/1848	1839	1848
Frederik 7.	1808	6/10/1808	1863	15/11/1863	1848	1863
Christian 9.	1818	08/04/1818	1906	29/01/1906	1863	1906
Frederik 8.	1843	03/06/1843	1912	14/05/1912	1906	1912
Christian 10.	1870	26/09/1870	1947	20/04/1947	1912	1947
Frederik 9.	1899	11/03/1899	1972	14/01/1972	1947	1972
Margrethe 2.	1940	16/04/1940	NA	NA	1972	NA

Note that we chose to include both birth_year and death_year as seperate columns as the death_date and birth_date were incomplete for several of the early Danish monarchs. Wikipedia was used (HTTPS://EN.WIKIPEDIA.ORG/WIKI/LIST_OF_DANISH_MONARCHS) in addition to the sources given in the task.

Some dates were impossible to find and were thus filled with *NA*.

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

Openrefine keeps the history of all changes that were made to the raw data and allows the user to undo changes. In this way, the raw data will always be safe if you need it.

(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?"

The interviews dataset contains the column *months_no_water* with values such as *["Aug"; "Sept"; "Oct"]* and *["Oct; Nov"]*. We transform the cells using GREL:

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

Which leaves us with values such as *Aug;Sept;Oct* and *Oct;Nov*. We then do a custom text facet where we split on the seperator ";"

```
value.split(";")
```

We get the following overview:

From this we note that the month of **October** is noted as the most water deprived/driest month followed by the month of **September**.

(4) 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, look at the HISCO classification "Historical International Standard of Classification of Occupations" on Dataverse if ambitious)

In OpenRefine, we firstly apply a text filter of the column *civiltilstand* to only show us the unmarried men and women (by writing "*ugift*"). We then split the column *erhverv* using regular expressions splitting on comma and the word "og" as many people have listed several occupations `,|\bog`. We then only work with the first listed occupation as we assume that that is their primary occupation.

A Text facet is used on the column *erhverv*. We then apply many different clustering algorithms. For instance, many occupations are somewhat close to Soldier (e.g., "Landsoldat", "Nationalsoldat", "Soldat") which have all been made into one group.

The result is the following 10 occupations: