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

<https://regex101.com/r/xhR7p5/1>

Kode: (\d+).(\d+).\s?(\d+)

Man skriver koden i ind i regex, og skriver i substitution: $3-$1-$2 for at få datoen til at stå i rigtig rækkefølge

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)**

**Voyant Stopword list->R Stopword list**

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

Kode: ([a-zæøå0-9éü’.?]+)(\n)?

Man skriver koden ind i regex, og skriver nede i substitution: "$1", for at få en voyant stopliste til at passe ind i R stopword List

**R Stopword list -> Voyant Stopword List**

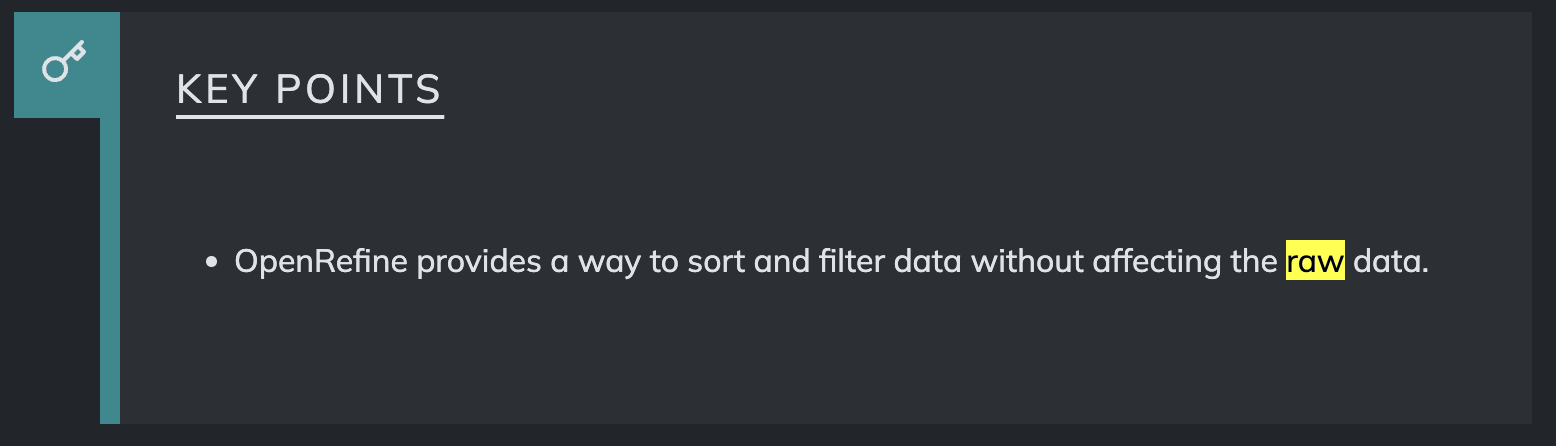
<https://regex101.com/r/Qz546I/1>

Kode: (["])([a-zæøå0-9.,éü’]+)(["])([,])?([ ])?

Man skriver koden ind i regex, og skriver nede i substitution: $2\n for at forvandle R stopword liste til en Voyant stopword liste

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

* Det gør den ikke



<https://datacarpentry.github.io/openrefine-socialsci/03-filter-sort.html>

1. 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?"



Jeg lavede en custom facet under month\_no\_water mod koden value.split(";"), for at få denne liste, derefter sorterede jeg efter “count” og ikke “name”

1. 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.)

Jeg sorterede så jeg kun fik ugifte mænd ind i arket, derefter satte jeg Cluster igang med at rette fejl i ordene, så som National Soldat og nationalsoldat bliver til National Soldat, så de bliver talt sammen i stedet for hver for sig

