Homework

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

**Result:** <https://regex101.com/r/YJGlUz/1>

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

**Results:**

Part 1: <https://regex101.com/r/IubwlB/1>

part 2: <https://regex101.com/r/cOpd4I/1>

1. In 250 words, answer the following question: "What are the basic principles for using spreadsheets for good data organisation?

Using spreadsheets consists of a few rules or principles. First at foremost you need to be consistent. Don’t use synonyms for the same word, keep it simple, e.g., ¨male¨ throughout your data instead of having ¨male¨, ¨man¨ etc. Spaces are tricky, so choose a simple and good name for your file. Use \_ instead of space because they can make programming harder. It is a good idea not to use a mismatch of different special characters such as (!\*(,)/) because they have a special meaning in programming language. To make the spreadsheet easier to work with, put the layout of the data into a shape of a rectangle, with one word per cell, this is to prevent missing data. Another very important principle is to create a data dictionary; it is a helpful tool that explains what all the variables are, and contains metadata, that gives information about the data you’re working with. Your spreadsheets are meant to be a collection of raw data, so even though you might be tempted to use different font colors, or highlights, it is best to keep it clean, this also means that your primary data file should not include calculations or graphs. Lastly make regular backups of your data, so that if something gets corrupted or an error might occur, you can go back and fix it.[[1]](#footnote-1)

1. Broman K.M. & Kara H. Woo 2018 Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10 [↑](#footnote-ref-1)