# Regex excercises

## Task 1 – Extracting and substituting dates to the format YYYY-MM-DD

Regular expression for extracting the dates: <https://regex101.com/r/qgn6z0/1>

Et billede, der indeholder tekst, skærmbillede, nummer/tal, Font/skrifttype

Automatisk genereret beskrivelse

The *\d+* is used to establish that we are looking for any digit. The *+* indicates that the preceding element, in this instance a digit, appears one or more times.

The *.* matches any special character, so the dot, comma or slash between the digits. \s matches any whitespace character so that would be the space in the dates like in the first date found, where there is a space between the comma and 1513.

The ? after \s? means that the whitespace character appears zero or one time.

The {4} means that the digits appears exactly 4 times

By using all of these in a regex we will be able to find all the dates and Regex101 will group date, month and year into different groups

Regular expression for substituting the dates into the date format: YYYY-MM-DD

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

Since regex have put date, month, and year into different groups it’s easy to find the different groups and change the order of the group, so that it will change the format to YYYY-MM-DD

$1 will capture the contents in group one, so in the first date the number 3 will be a part of group 1. 27 will be a part of group 2 and 1513 is a part of group 3. After that it is simple to change the order, so that it will change to the format YYYY-MM-DD

## Task 2 – Converting the Danish stopwordlist from Voyant to R and the from R to Voyant

**Voyant to R** - <https://regex101.com/r/BsYqKI/1>

Et billede, der indeholder tekst, skærmbillede, nummer/tal, Font/skrifttype

Automatisk genereret beskrivelse

\n is used to find all of the linebreaks. After finding the linebreaks you can substitute the linebreaks into commas and therefore one continuous stopwordlist, which is ideal when working in R. You do that by using *quote on quote, “,”*

**R to Voyant –** [**https://regex101.com/r/f6cT01/1**](https://regex101.com/r/f6cT01/1)

**Et billede, der indeholder tekst, skærmbillede, software, nummer/tal

Automatisk genereret beskrivelse**

To change the stopwordlist from a stopwordlist suitable for R into af stopwordlist suited for Voyant, all you must do is reverse the regex from before.

## What are the basic principles for using spreadsheets for good data organisation?

When using spreadsheets there is a lot of different methods and principles for how to keep the spreadsheet and data tidy and organized. Programs like excel contains certain bugs and some formats can be different depending on the program that’s used. One important rule when working with spreadsheets is to be consistent when it comes to categories and abbreviations. The same goes for dates, as shown above in task 1, you should always use the same date format. When being consistent, it also makes it easier to automate the process, since the categories and abbreviations have the same names throughout the entire dataset.

Another important thing is to create backups, so it is always possible to find a version of the spreadsheet in case something should happen to the spreadsheet.

Another way of organizing your data is to keep your spreadsheet focused on strictly data itself and then create other files for your analyzation, calculation, or visualization of the data. Using good names for your files or backups are also crucial in organizing your data, it will make the spreadsheet more navigable for the person working and creating the spreadsheet, as well as the future readers of the spreadsheet.

In general the most important aspects when working in spreadsheets is to make the spreadsheet easier to automate and to make it more navigable.