# Homework W36: Regular expressions

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?**

The regular expression used here to find all dates:

**\d{1,2}.\d{1,2}.\s?\d{4}**

The **\d** class matches any digit, while **{1,2}** specifies a range, so that only numbers that are 1 to 2 digits in length are found. The metacharacter **.** matches any character. Any white space character is found using the **\s** class, while the **?** metacharacter matches, when the preceding element appears zero or one time. Lastly, **\d{4}** matches any sequence of four-digit numbers.

By combining these characters, all six dates that appear in the text extract in question are found.

By using parentheses **( )**, the characters signifying month, day and year are captured into separate groups:

**(\d{1,2}).(\d{1,2}).\s?(\d{4})**

Each of the captured groups are automatically numbered, so that Capturing Group 1 correlates to month, Capturing Group 2 correlates to day and Capturing Group 3 correlates to year.

By using the in-built substitution function, each of the capturing groups are selected in the taskbar using **$**. To convert into YYYY-MM-DD, the capturing groups are written in an order, corresponding to the desired format, separated by a hyphen:

**$3-$1-$2**

Link to regular expression: <https://regex101.com/r/xtFpSe/1>

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)**

***Stopword list to R***

The stopword list needs to be a converted from a list separated by line breaks into one, where each of the words are in quotation marks and separated by commas (and a white space character).

The regular expression **\n** matches any line break (newline character). Using the substitution function, the line breaks are then replaced by a comma and a white space character with quotation marks on either side (i.e. **“, ”**).

Most of the work can be done by implementing the method above. However, the first and last words need to be manually changed before use, as they are missing a single quotation mark before and after respectively.

Link to regular expression: <https://regex101.com/r/UGeNZN/1>

***Stopword list to Voyant***

This stopword list needs to be converted from a list, where each individual word is in quotation marks and is separated by a comma and a white space character, into a list, where the words are separated by line breaks.

The following regular expression is used to match quotation marks and interpunctuation: **“, ”**. Using the substitution function, this sequence of characters is then replaced by a line break **\n**.

As with the case above, this method is not complete – there are still quotation marks at the beginning and the end of the list that need to be manually removed before being put to use in Voyant.

Link to regular expression: <https://regex101.com/r/NZ8CJb/1>

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

Consistency is a key principle of good data organization. This is the case in all aspects of data organization, including naming as well as layout and format.

When naming it is important to use a system that is concise but meaningful and then stick to it, e.g., by using the YYYY-MM-DD date format. This will make the data sheet more manageable for one’s future self as well as anyone else who may want to analyze the data. Inconsistencies in naming can hinder machine readability as well.

Another basic principle of good data organization is to fill in all cells. “NA” or any other common identifier should be used to signify missing values. It is also important to ensure that each cell only contains one piece of information. Annotations should be written in a separate column. Likewise, it is recommended not to use font color or high light to signal information about the data.

When tidying data, layout is another factor to keep in mind; one single rectangle is the most appropriate shape.

It is also practical to create a separate ReadMe file as well as a file containing important metadata. The file with the raw data should not have any other information than the data itself, i.e., no calculations or graphs either. It is also recommended to use data validation as to avoid errors.

Making backups of one’s data and write-protecting it is another key principle of data management. Likewise, it is a good idea to copy and save the data in a plain text file.