**Regular expression**

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 ?**

To chance the dates so they all are in the same form, I start by separating the days, months, and years. This is done by writing

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

When I have all the different digits highlighted, I am then capable of changing the order in which the numbers are written. I will chance the dates to the format YYYY-MM-DD, this is done by using the expression.

$3-$1-$2

<https://regex101.com/r/QS2lfK/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)**

The aim is to convert the voyant list to a stopwordlist. This is done by first finding the separate lines without interpunction. For this I used the regular expression

(\n)

When I have found the lines, I can then use the expression

“,”

To get the words separated by a comma.

<https://regex101.com/r/2aBwol/1>

To convert a stopwordlist in to a voyant. The first step is to isolate all the commas. This is done by using the expression

“, “

The expression contains a space between the comma and the quotation mark. Then to get a list where each word is on a separated line, I use the expression.

\n

<https://regex101.com/r/IMUE7e/1>

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

Some of the basic principles for using spreadsheets for good data organization are to be consistent, to choose good names, to create a data dictionary, make backups and data validation. These are just some off the principles mentioned in the article “Data organization in spreadsheets” by Karl W. Broman & Kara H. Woo. The first principle to be consistent means, that throughout the data file you must be consistent when it comes to for example names or phrases, the names can’t be spelled in five different ways through the file. When it comes to choosing a name, the name should be short and meaningful. It is also a good idea to use the global standard when writing a date. By doing so it becomes easier for others across the world to read the file when they already know the format. Another principle that makes it easier to read and analyze the data file is to create a data dictionary. A data dictionary helps explaining the different variables there is used in the data file. It is a good principle to always make backups and make them regular and in different places. Another basic principle is to keep a copy of the data file in plain text format. In the raw data file should there be no calculations. By doing calculation in the raw data file, you risk changing some of the original data or you might make a mistake and delete some of the data. But most importantly is to be consistent through the data file.