**Work on regular expressions work on regular expressions**

**Done by Trine**

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?***

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

Et billede, der indeholder tekst, elektronik, skærmbillede, software

Automatisk genereret beskrivelse

In the test string:

First, I insert the text and then I use “(/d+) to match every single digit which stands for either year or month in the text. Then I use “.”, a regular expecion which matches all characters and then “(/d+)” to catch the month or year again (I use it twice since sometimes the year stands before the month and vice visa). Then I use “\s?”, since the original text includes special characters between the dates, and we want to delete those. I finely write “(\d {4})” to capture every digit 4 times and there are 4 digits in a year.

In the substitution:

I use “$3” to capture my third capturing group (the year), in that way the text will transform and write it as the first thing. Then I tab “$1” to capture the first group (the month), in that way it will stand as the second. At last I tab “$2” to include the second capturing group (the date) e voilá

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)***

<https://regex101.com/r/7TYYyV/1>

Et billede, der indeholder skærmbillede, tekst, software, Webside

Automatisk genereret beskrivelse

Test sting:

First, I insert the text and then I write “”, •”” to delete the commas and the spaces between the words.

Substitution:

I write \n to get the words on a line.

Et billede, der indeholder tekst, software, Webside, Computerikon

Automatisk genereret beskrivelse

<https://regex101.com/r/Cmt8cu/1>

Test string:

First, I insert the second stop wordlist and write “\n” to delete the spaces between the words.

Substitution:

I insert “,” to make a comma between the words.

Et billede, der indeholder skærmbillede, tekst, software, Webside

Automatisk genereret beskrivelse

<https://regex101.com/r/kk9G3P/1>

To do it the other way around I just twist the regular expressions.

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

* You must write it in a rectangular shape to keep it tidy and nice.
  + Make rows for the subjects (please only use one row for those) and for the variables.
* No empty spaces, instead write “\_” or “NA.”
* Make different documents for different sets of data.
* Make your datasets assessable to other people.
* Create short and specific names to your documents.
* Do not make calculations in your datasets, excel doesn’t like it and it might mess with your raw data.
* Make backups. Consider saving the data in plain text files.
* Be consistent in whatever you do. If you make up a short name for something, then stick to it
* It is a good idea to write dates as following: YYYY.MM.DD, since It is the global standard <3
* Consider adding a “note column” and remember to transform the area to a text area. Do not mix data and text.
* Maybe have a different file which explains what the variables are.
* You could also create a “ReadMe file” which explains the data set, to make it more accessible to others, a so-called data dictionary.
* Do not use highlighting of words in your dataset, even though it is tempting. It might be difficult for others and your future self to understand the function of the highlight.

The most important thing to do: Be consistent!