**Homework 1 – 12/9 2023**

**Question 1**

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 ?*

We have used the regular expression : \d+.\d+.\s?\d{4}

/d matches any single digit between 0-9. + matches the preceding element (d) one or more times. The ‘.’ (without parentheses) matches any character. s matches any space, tab or newline. ? matches when a character appears 0 or one time. With these metacharacters we highlight every date in the text, so we know what to work with.

Se following regex-link : <https://regex101.com/r/mZNgej/1>

By putting parentheses () around each group we highlight the date, the month, and the year separately : (\d+).(\d+).\s?(\d{4}).

See the following regex link : <https://regex101.com/r/OxFoDW/1>

When we have to change and put it in the same following format YYYY-MM-DD, we first change the window to the substitution window. Here we insert the following replacement value : $3-$1-$2. The anchors $ captures the group that is identified by a number in the replacement string, where the number is the decimal value. The – marks that sign we want to separate our values with. Hereby we get the form we want for all dates.

See the following regex link : <https://regex101.com/r/VenbQq/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)*).*

Here the stopword list from Voyant is copied into the regex in the substitution window. Here we indicate \n in the regular expression area to mark the newlines between the characters. To convert to a stopword list for R, the replacement value “, “ is put to create the form we want the list to be in.

Se the following regex link : <https://regex101.com/r/sAHv4C/1>

The first and last word in the list is not entirely surrounded by the “” sign, but it is a close attempt.

*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)*

Here it is the other way around. The R stopwordlist is copied into the regex and the “” surrounding each word is highlighted by tapping the regular expression “, “ in.   
In substitution the replacement value \n is put in and the list changes to Voyant format.

Se the following regex link : <https://regex101.com/r/LIJdBT/1>

There is a “ at the first and last word of the list, but still a close attempt.

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

Spreadsheets are an important part of researching to organize, analyze and visualize often complex data. But they have a history of containing errors and being difficult to manage. Therefore, a set of practical rules to avoid these errors as well as to create a better workflow and research, processable to computers and to share with others, are much welcomed.

* 1. Rule number on is to be consistent with everything you do in a spreadsheet while organizing by using consistent fixed codes, names, layout, format etc.
  2. Choose good names for things – short and meaningful are key words here.
  3. Remember to use the global standard for writing dates YYYY-MM-DD
  4. Don’t leave any missing cells but fill them in instead. Use a common code for missing data to make it clear that it is missing.
  5. Each cell in the spreadsheet should only contain one piece of data to show the exact information clearly.
  6. The best layout for the data in a spreadsheet is a rectangle. The rows represent the subjects while the columns correspond to variables. If necessary, make several rectangles.
  7. Create a data dictionary. It explains what all the variables represent and includes information about the data with all the details that make sense for the dataset.
  8. The spreadsheet should only contain the data and nothing else.
  9. Also, do not highlight any part of the data.
  10. Making back-ups is important. Keep all the old versions for later use. Use the data validation in Excel to avoid data entry errors, and last but not least, save the data in plain text files like CSV. Easy to open and work with.   
        
      But all in all, the key is consistency.

1. *Challenge (OPTIONAL)!Can you find all the instances of 'Dis Manibus' invocation in the EDH inscriptions in*[*https://bit.ly/regexexercise5*](https://bit.ly/regexexercise5)*? Beware of the six possible canonical versions of the Dis Manibus formula!*

We have copied all EDH inscriptions into the regex. Our focus was that the six following are the canonical versions of the Dis Manibus formular : D M, D M S, Dis Manibus, Diis Manibus, Dis Manibus Sacrum, Diis Manibus Sacrum.

Starting the regular expression to capture all canonical versions, it is important to start with the longest version of the word Diis Manibus Sacrum to be able to catch the shorter versions as well. Therefore, the whole regular expression is :

\b[Dd]i+s [Mm]anibus [Ss]acrum\b|\b[Dd]i+s [Mm]anibus\b|\b[Dd] [Mm] [Ss]\b|\b[Dd] [Mm]\b

The \b marks a word boundary. The [] take into account both uppercase and lower-case letters. They are repeated for each version. The | marks or and separates the words searched after. The + matches the preceding element one or more times.

We found 23.786 instances of the Manibus invocation.   
There might be some left out, but this seems that it could be a good deal of them.

To see for yourself go to regex101.com and copy the data and the regular expression in. The data set is way too big to be saved and copied in a link.