Homework 1- Regular Expressions

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?**

You can find my answer by following the link:

https://regex101.com/r/n0tgYq/1

And here are the text piece: *"Juan Ponce de León sights Florida for the first time, on 1513-3-27. Giovanni da Verrazzano explored the Atlantic coast of North America under French employ, on 1524-4-17. The Roanoke Colony was found deserted, on 1590-8-15. John Smith founded the Jamestown settlement, on 1607-5-14. The Dutch laid claim to the territories of New Netherland, on 1614-11-11. The Massachusetts Bay Colony founded, on 1629-3-4"*

I used the regular expression (\d{1,2}).(\d{1,2}).(\s?\d{4}) to make 3 different groups which matches all the different date format in the text-piece, the "." Represent the different signs between the numbers. Afterwards I use the substitution function, changing the order of the groups I made in the regular expression: month, date and year ($3-$1-$2), and then separate them by the same sign "-" making the dates uniform.

2. **Write a regular expression to convert the stop wordlist (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)**

Here you'll see the Voyant list converted to stop wordlist where the words are separated only by commas: https://regex101.com/r/fdXFKx/1 I used the expression \n to mark all the line shifts, and then substituted them with comma using the substitution function and \,

Here you'll see the second wordlist converted into a Voyant list: https://regex101.com/r/cr1alH/1 I used the regular expression \", "|[",] to catch the signs separating the words: the ", " or the ",

Then I used the substitution function and typed: \n, and then the words were separated by line shift

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

To effectively use spreadsheets for good data organization, several basic principles should be followed. First of all you have to make sure, that you have good quality data. It must be accurate, valid, and often representative. Your dataset also needs a certain degree of transparency, in that way you secure that people who may use the data, knows that your data is valid.

Once you have quality and transparent data, the next step is to organize it effectively within the spreadsheet. Spreadsheets can be powerful tools, but it need a clear and consistent structure. Hadley Wickham formulated some key principles for organizing a spreadsheet, named "tidy date", which include that each variable should have its own column and each observations its own row. Furthermore, all the values should have its own cell.

But to secure a good structure you may need good headings or descriptions that will make It easier to navigate in the data set. But don't write down the descriptions within the dataset, make a separate document. Also, when working with the raw data, you should do it in a copy of the original spreadsheet, in that way you will avoid messing with the data.

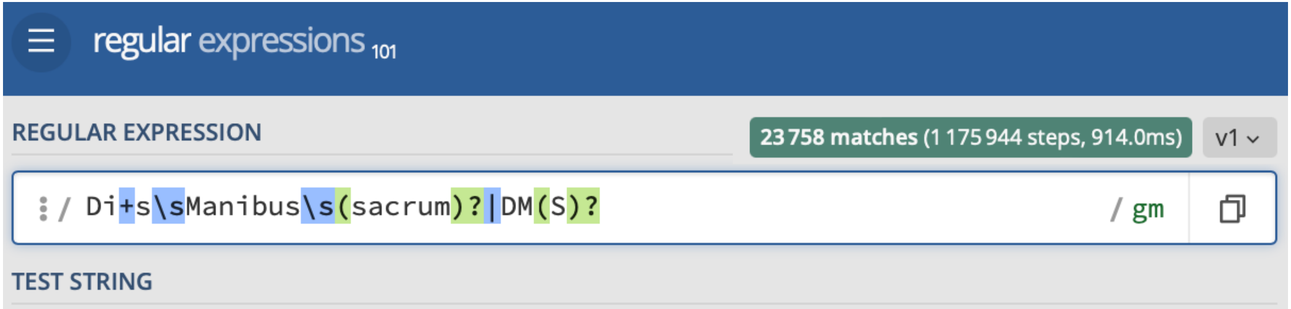
Another thing you may have to make sure, is that the data in each column follows the same format, it can help you to prevent mistakes and spare you some time. To avoid confusing the computer you should also avoid creating multiple data tables within one spreadsheet.

Butt all the above-mentioned points doesn't matter if you, or others, cant gain access to the spreadsheet, so what's more important is that the data is findable and the spreadsheet is easy to access. You will also have to make sure that the raw data in the spreadsheet is reusable, so others can recreate and check your results.

In conclusion, to use spreadsheets effectively for data organization, start with high-quality data, maintain a clear and consistent structure, and be sure that it's easy to find and reuse the dataset. Once you got It covered you will have a spreadsheet with good data organisation.

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

One of the regular expression I've used:



23 758 matches (1 175 944 steps, 914.0ms)

```
/ Di+s\sManibus\s(sacrum)?|DM(S)? / gm
```

But due to the data size, I'm unable to provide the link.