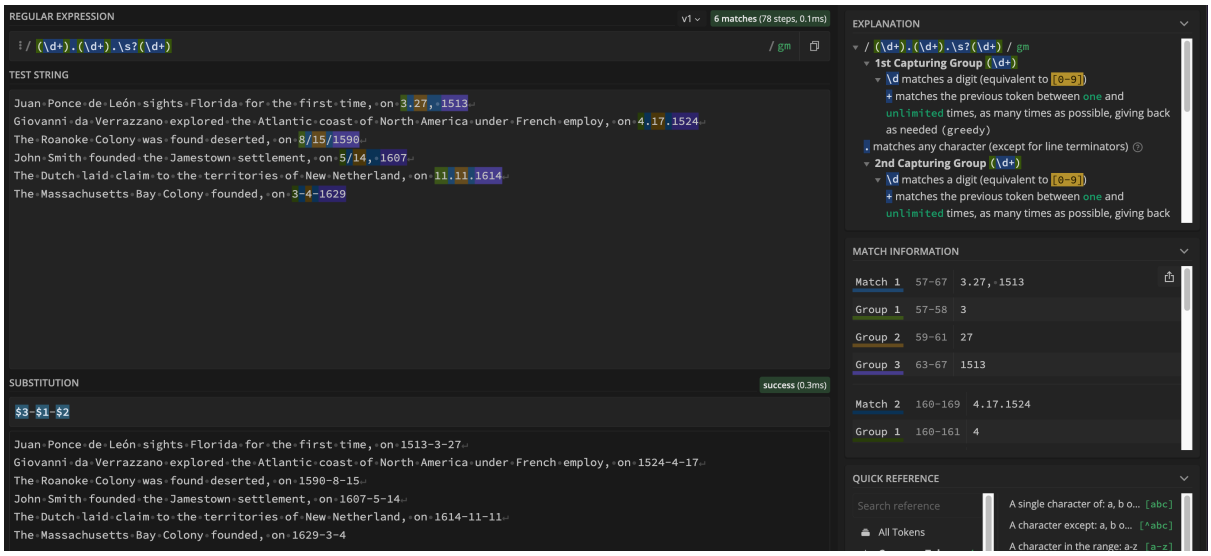
Anton Drasbæk, AU682983

1: W35: Regular Expressions and spreadsheets

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 ?

Link to my solution: https://regex101.com/r/X0NOyU/1

Screenshot of my solution:

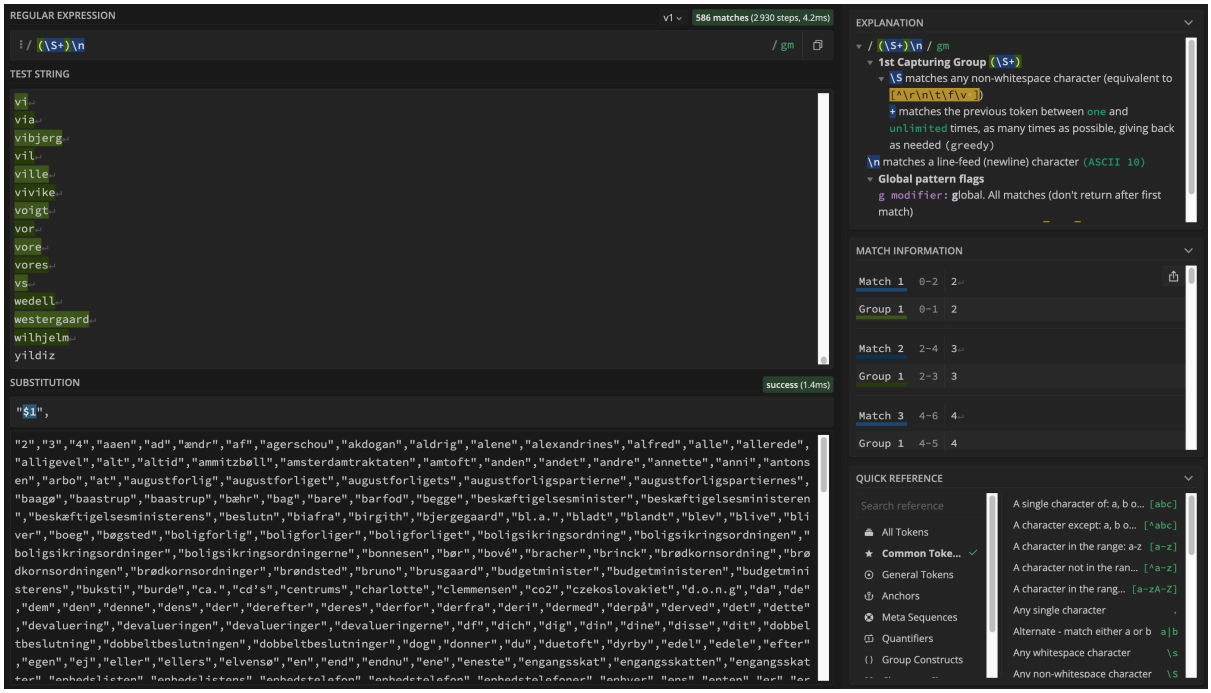


2. Write a regular expression to convert the stopwordlist (list of most frequent Danish words) from Voyant in http://bit.ly/regexexercise3 into a neat stopword list for R (which comprises "words" separated by commas, such as http://bit.ly/regexexercise4). Then take the stopwordlist from R http://bit.ly/regexexercise4 and convert it into a Voyant list (words on separate line without interpunction)

Link to RegEx from Voyant format to neat R-format: https://regex101.com/r/maJ8gI/1

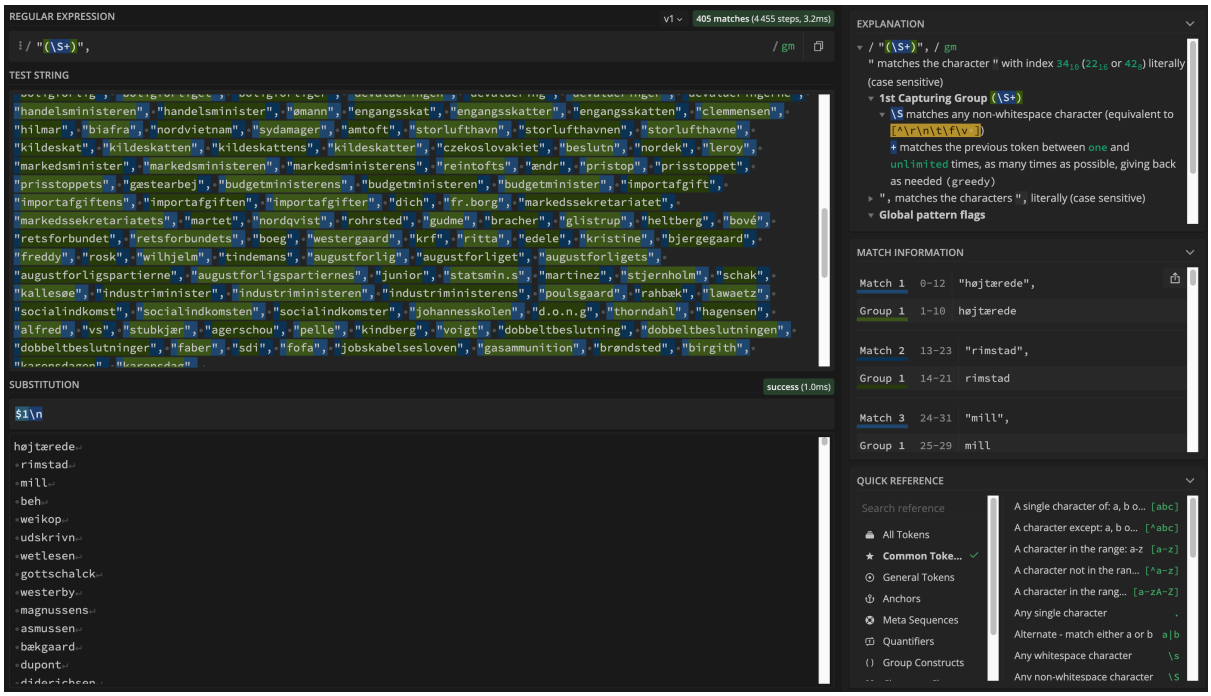
Screenshot of RegEx from Voyant format to neat R-format:

Anton Drasbæk, AU682983



Link to RegEx R-format to Voyant list: https://regex101.com/r/5GCNUN/1

Screenshot of RegEx R-format to Voyant list:



Anton Drasbæk, AU682983

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

There are a few basic principles for using spreadsheets for good data organisation. Firstly, data should be structured in such a way that each column indicates one variable (e.g. age, not age and height) and each row indicates one observation of each variable (e.g. one household). Secondly, data should be entered into the spreadsheet in a consistent and accurate manner. This means using the same format for the data and ensuring that all data is entered correctly. This would also include using consistent formats for missing values (e.g. not using -99, -999, NA and empty cells). The principle for consistent format also holds for other variables such as not using 1, YES, yes, Y, etc. all to indicate “yes” for an observation of a variable. Moreover, the data should be setup so that it is exportable for a CSV or similar structure. This is necessary in order to ensure that it can be worked with outside of the spreadsheet. This is achievable by not including several tables in one spreadsheet, using multiple tabs if not necessary, and abstaining from using colors to indicate a variable as that does not convert well when importing into other software such as R. Finally, variables should be named in a way that is succinct yet easily understandable. Especially if no metadata is included, the variables should be named in a manner such that third parties easily can understand what observations of that variable means.