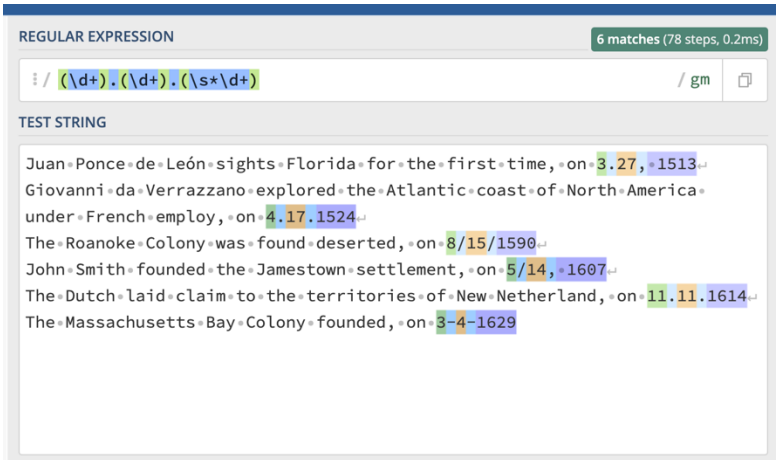
Ida Marie Byskov Jensen

1. What regular expressions do you use to extract all the dates in this blurb: http://bit.ly/regexexercise2and to put them into the following format YYYY-MM-DD ?

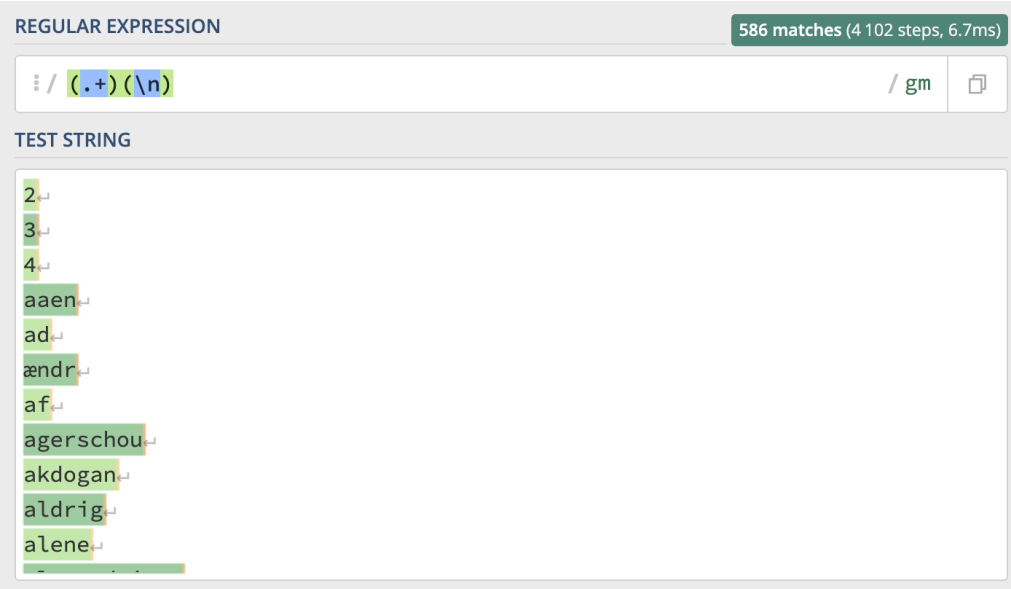


Explanation: I group them by doing the parentheses. The \d match any digit and then the + is because it can appear one or more times. This is the same for the next part. For the last part the \s matches any space, tab or newline. The * is so that the preceding element (\s) match zero or more times, as in this case is where the space can both be present or not in the text.

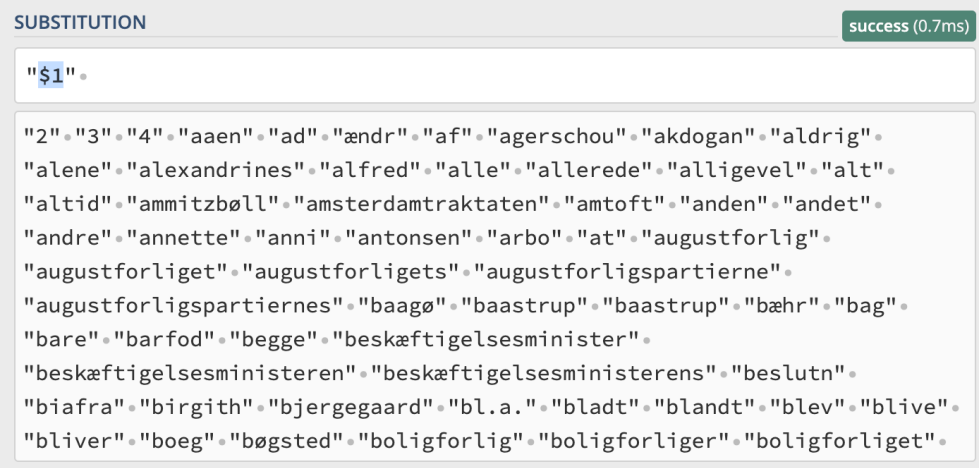
In the above part, by dividing them into groups, this makes it possible to substitute them and change the order. This is done by opening the substitution window. The $ sign, are used to find each of the above groups.

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)

First part is to convert the stopwordlist from Voyant into a stopwordlist for R;

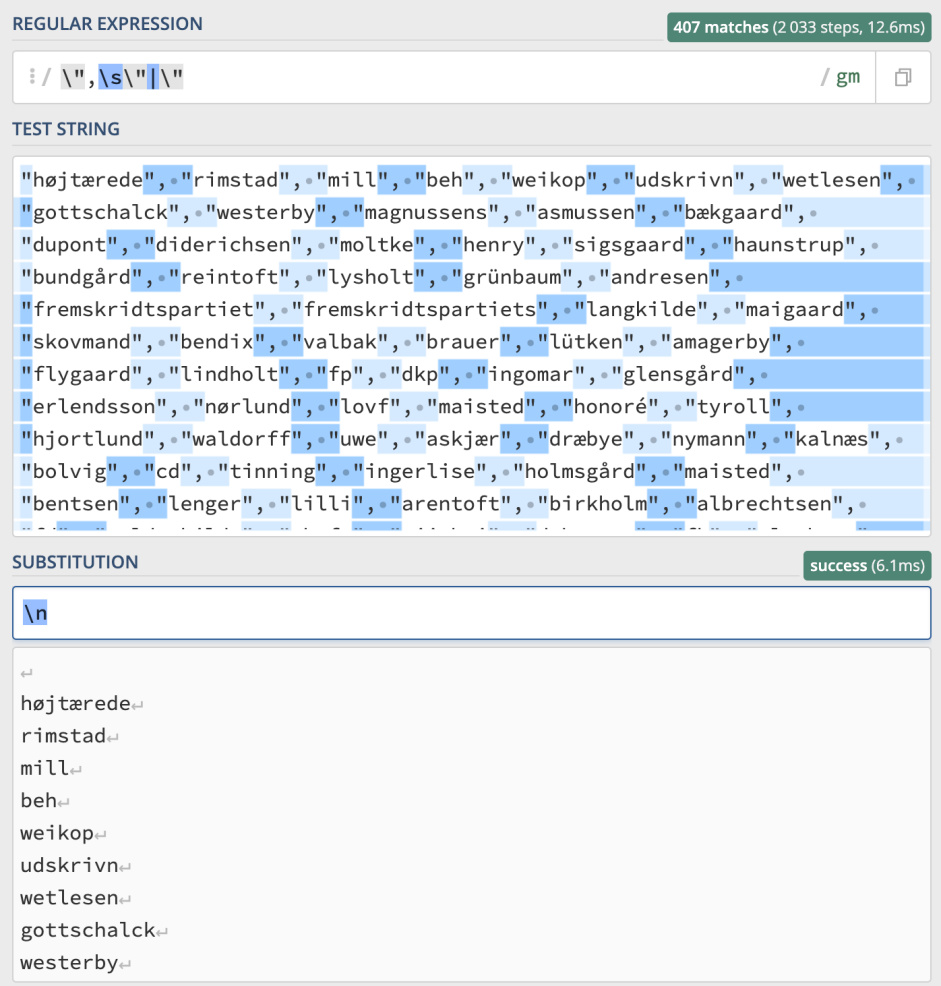


Using the parentheses to group them. To find both digits and characters I use . (matches any character) and + as it can appear one or more times. Then I use \n to match line breaks.



Each word is now marked as group 1. By adding quotation signs around the group, and making a space in the end, it turns out as the same format as the R stopwordlist.

To convert the R stopwordlist back to Voyant, \" match all quotation signs. Then match the comma and the \s matches any space, tab or newline. Then the | mean ‘or’, which means that the whole regular expression should match on either the right or left side of this sign. \" is to catch the fist and last quotation signs.



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

Data can be organized in many ways, one is spreadsheets, this can be XLS or XLSX (and more) as known, for example, from Excel. Excel spreadsheets are a widely used software tool for data entry, storage, analysis, and visualization.

Before doing any organization and overwriting, always make sure to have a copy of the raw data as it is originally generated. This is essential for the opportunity to rerun the analyses. It's also practical in case of analytical mishaps, and for experimenting without fear. A way of avoiding overwriting the original file, could be to set the file permission to 'read only'.

To organize these spreadsheets, the most important thing, is to be consistent. Write data the same way in every cell and don't leave any empty cells. In case of missing data write 'NA', which is a code used in most programming languages to indicate that data are 'Not Available'. Be very aware of the naming, of both files and columns.

- Files should be named so that they are easy to find and match. A good filename for meta data would be '2015-04-denver-csv'. This naming makes it easy to match both by date and location.

- Rename columns that are instructible and artificial codes. An example could be to rename name1 and name2 to personal_name and family_name.

Another important principle is to only put one thing in a cell. Therefore you should not merge cells either. Also use the rectangular format, where the first row (the column names) is variable names, and the next rows are values to the corresponding variables. The last important principle is to make backups and save it often.

source references

Wilson et al. "Good enough practices in scientific computing"