Week 1 - Regex/Tables

Emil Bæk Mogensen

2023-09-14

Packages

```
library(tidyverse)
```

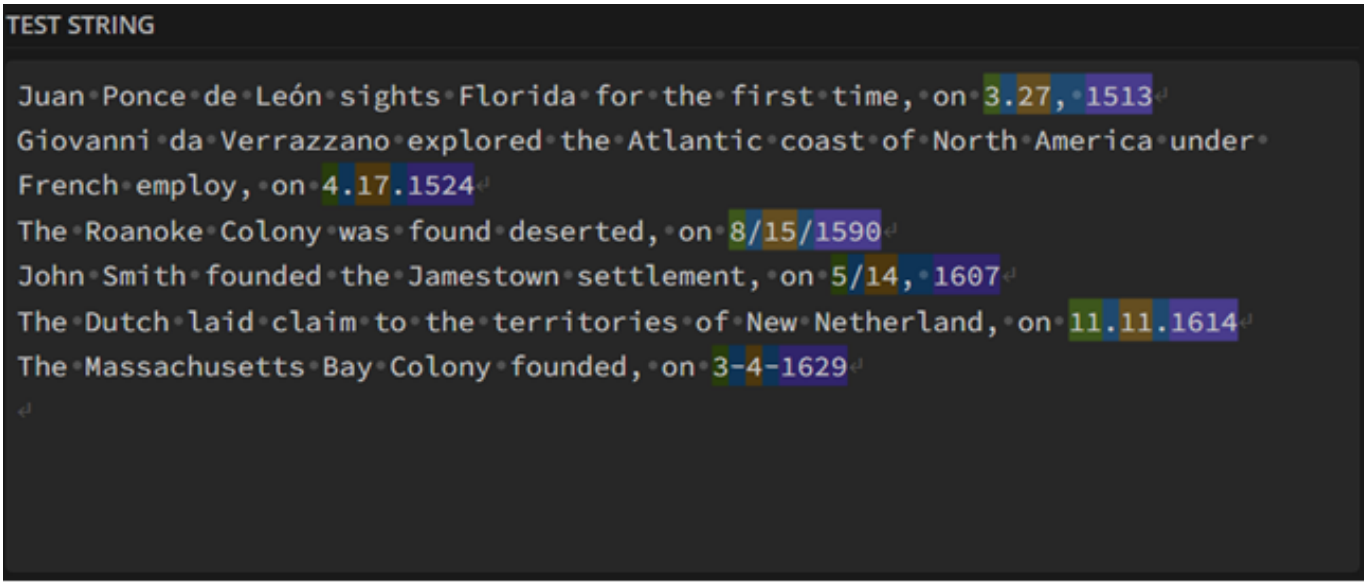
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

1a) What regular expressions do you use to extract all the dates and to put them into the following format YYYY-MM-DD?

```
#Regex: \d{1,2}.\d{1,2}.*?\d{4}
```

To make them the same format grouping is needed of the above explanation. The regex will then look like the following

```
#Regex: (\d{1,2}).(\d{1,2}).*?(\d{4})
```



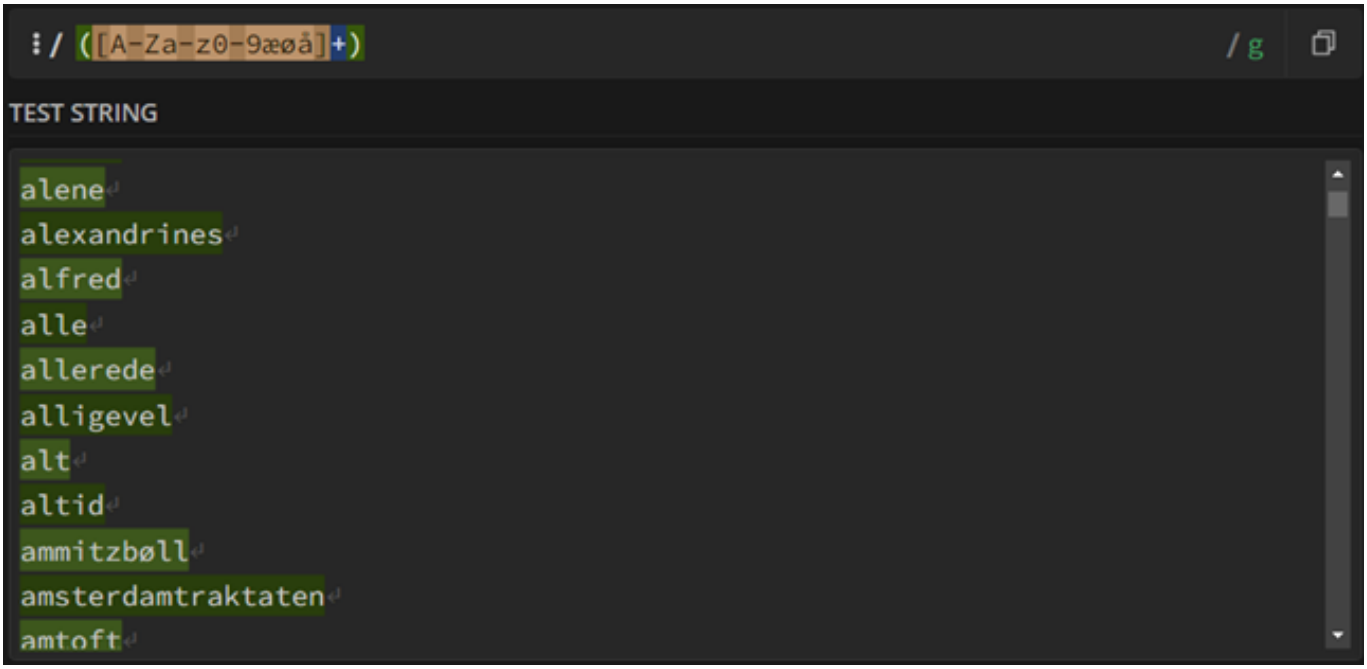
The substitution pattern formats the text into the same format:

```
#Substitute: $3-$2-$1
```

2a) Write a regular expression to convert the stopwordlist (list of most frequent Danish words) from Voyant into a neat stopword list for R (which comprises "words" separated by commas.

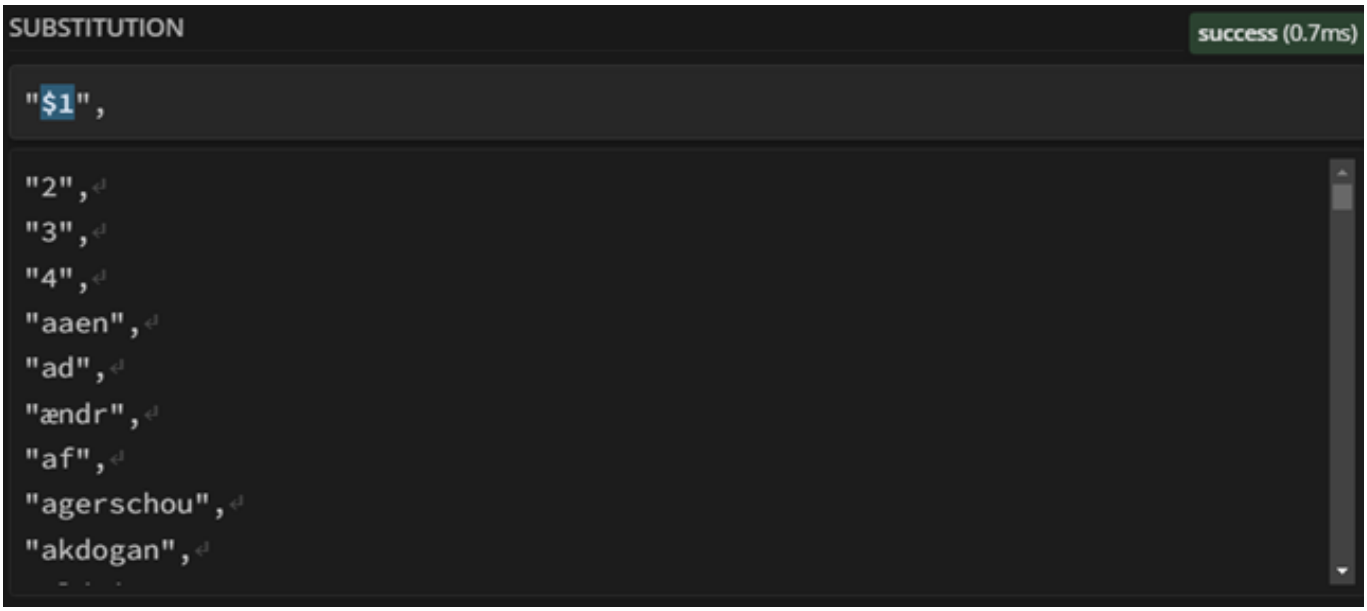
The following substitution makes the stopword list for R as in words separated by commas.

```
# Regex: ([A-Za-z0-9æøå]+)
```



The following substitution makes the stopword list for R as in words separated by commas.

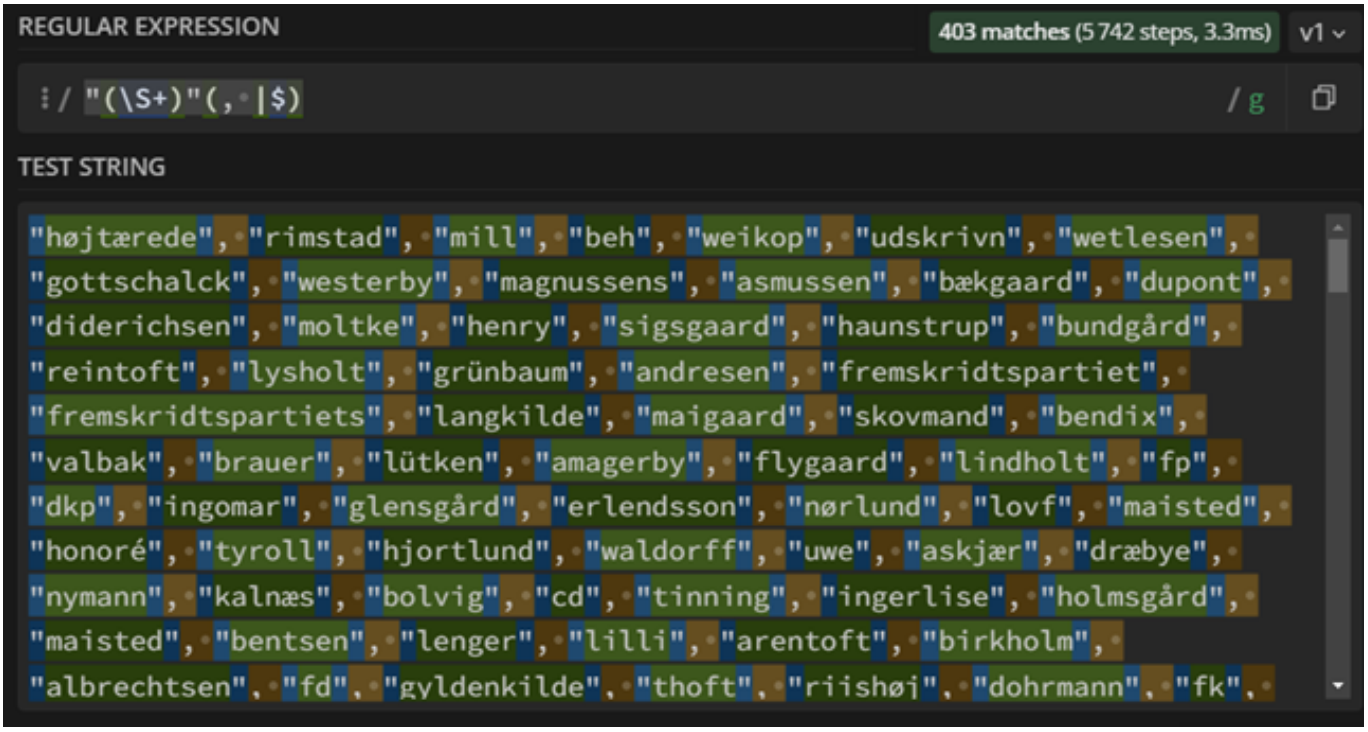
```
# Substitute: "$1",
```



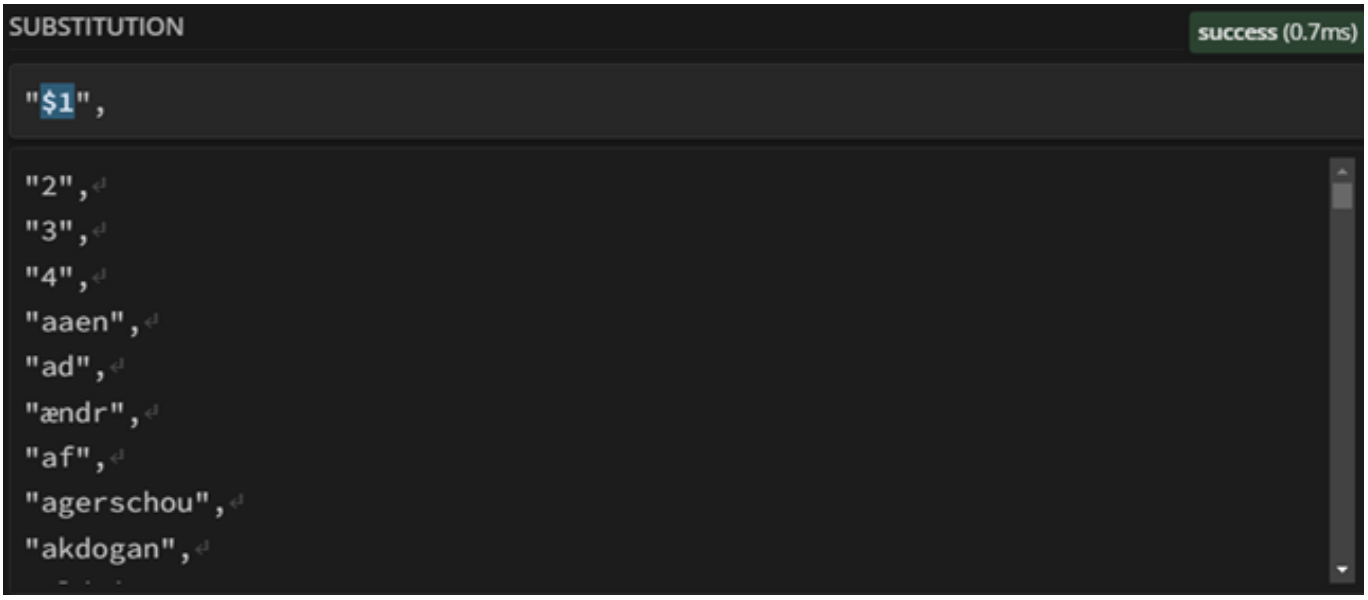
2b) **Then take the stopwordlist and convert it into a Voyant list of words on separate line without interpunction) To convert the R stopwordlist into Voyant list:**

To convert the R stopwordlist into a Voyant list I used the expression

```
# Regex:  "(\S+)"(, |$)
```




```
# Substitute: $1\n
```



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

```
# When working with big datas in programs such as excel, it is important to do all tables in different
#In tidy data each variable forms a column, each observation forms a row, and each observational unit f
```

4) **Challenge (OPTIONAL)!Can you find all the instances of 'Dis Manibus' invocation? Beware of the six possible canonical versions of the Dis Manibus formula!** The following regex should find all the six instances of the Dis Manibus invocation:

```
# \b(?:D\s?M(?:\sS)?|Di[si]{1,2}\sM[ai]{1,2}nibus(?:\sSacrum)?)\b
```