

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}
```

TEST STRING

```
Juan Ponce de León sights Florida for the first time, on 3.27.1513
Giovanni da Verrazzano explored the Atlantic coast of North America under
French employ, on 4.17.1524
The Roanoke Colony was found deserted, on 8/15/1590
John Smith founded the Jamestown settlement, on 5/14, 1607
The Dutch laid claim to the territories of New Netherland, on 11.11.1614
The Massachusetts Bay Colony founded, on 3-4-1629
```

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})
```

TEST STRING

```
Juan Ponce de León sights Florida for the first time, on 3.27.1513
Giovanni da Verrazzano explored the Atlantic coast of North America under
French employ, on 4.17.1524
The Roanoke Colony was found deserted, on 8/15/1590
John Smith founded the Jamestown settlement, on 5/14, 1607
The Dutch laid claim to the territories of New Netherland, on 11.11.1614
The Massachusetts Bay Colony founded, on 3-4-1629
```

The substitution pattern formats the text into the same format:

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

SUBSTITUTION

success (0.3ms)

```
$3-$1-$2
```

```
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
```

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

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

```
# Regex: ([A-Za-z0-9æøå]+)
```

```
:/ ([A-Za-z0-9æøå]+) / g
```

TEST STRING

```
alene
alexandrines
alfred
alle
allerede
alligevel
alt
altid
ammitzbøll
amsterdamtraktaten
amtoft
```

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

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

```
SUBSTITUTION success (0.7ms)
```

```
"$1",
"2",
"3",
"4",
"aaen",
"ad",
"ændr",
"af",
"agerschou",
"akdogan",
- - -
```

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+)"(, |$)
```

REGULAR EXPRESSION 403 matches (5 742 steps, 3.3ms) v1 ▾

```
:/ "(\\S+)"(,|\\$)
```

TEST STRING

```
"højtærede", "rimstad", "mill", "beh", "weikop", "udskrivn", "wetlesen",
"gottschalck", "westerby", "magnussens", "asmussen", "bækgaard", "dupont",
"diderichsen", "moltke", "henry", "sigsgaard", "haunstrup", "bundgård",
"reintoft", "lysholt", "grünbaum", "andresen", "fremskridtspartiet",
"fremskridtspartiets", "langkilde", "maigaard", "skovmand", "bendix",
"valbak", "brauer", "lütken", "amagerby", "flygaard", "lindholt", "fp",
"dkp", "ingomar", "glensgård", "erlendsson", "nørlund", "lovf", "maisted",
"honoré", "tyroll", "hjortlund", "waldorff", "uwe", "askjær", "dræbye",
"nymann", "kalnæs", "bolvig", "cd", "tinning", "ingerlise", "holmsgård",
"maisted", "bentsen", "lenger", "lilli", "arentoft", "birkholm",
"albrechtsen", "fd", "gyldenkilde", "thoft", "riishøj", "dohrmann", "fk",
```

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

SUBSTITUTION success (0.7ms)

```
"$1",
"2",
"3",
"4",
"aaen",
"ad",
"ændr",
"af",
"agerschou",
"akdogan",
```

- 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
```