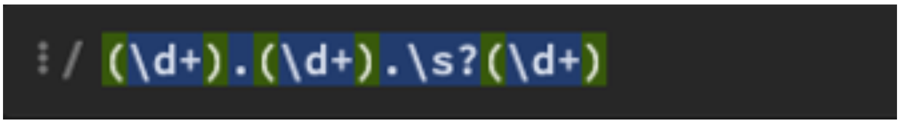
Homework 1

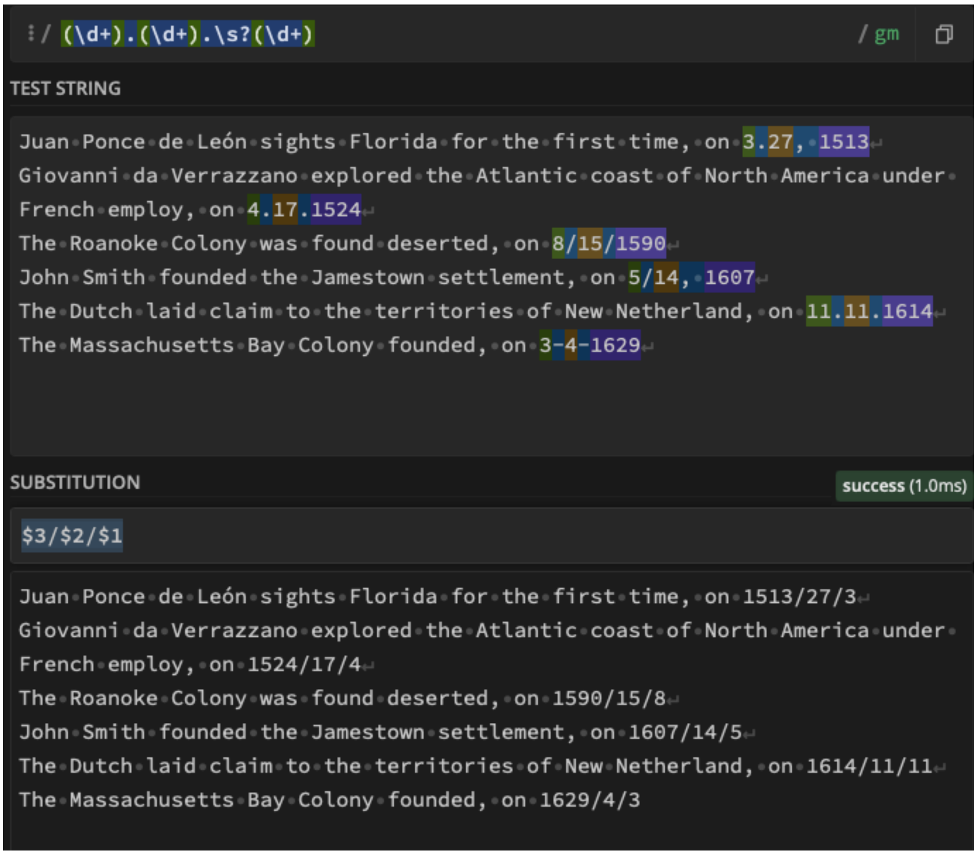
Task 1

To extract all the dates in the blurb I type the regular expression:



```
/ (\d+).(\d+).\s?(\d+)
```

First, I match any single digit which appear one or more times (\d+). I write this to times to catch both the day and month of the dates. Then I match any space, tab or newline which occurs zero or one time (\s?). Finally, I match any single digit which appear one or more times to catch the year of the date (\d+). The full stops match any character, and I use them in my regular expression, because a date often contains either full stops, slashes, or hyphens. I surround the dates with parentheses. The first parentheses can be represented by $1, the second by $2 and the third by $3. I click on "Substitution" in my and type $3/$2/$1. This changes the order of the dates into YYYY-MM-DD.



```
/ (\d+).(\d+).\s?(\d+)                                        / gm
```

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
```

SUBSTITUTION

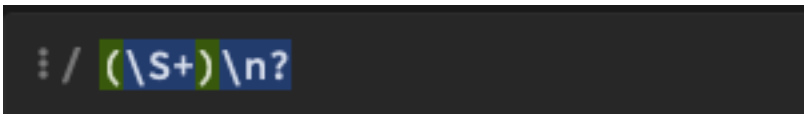
success (1.0ms)

```
$3/$2/$1
```

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

Task 2

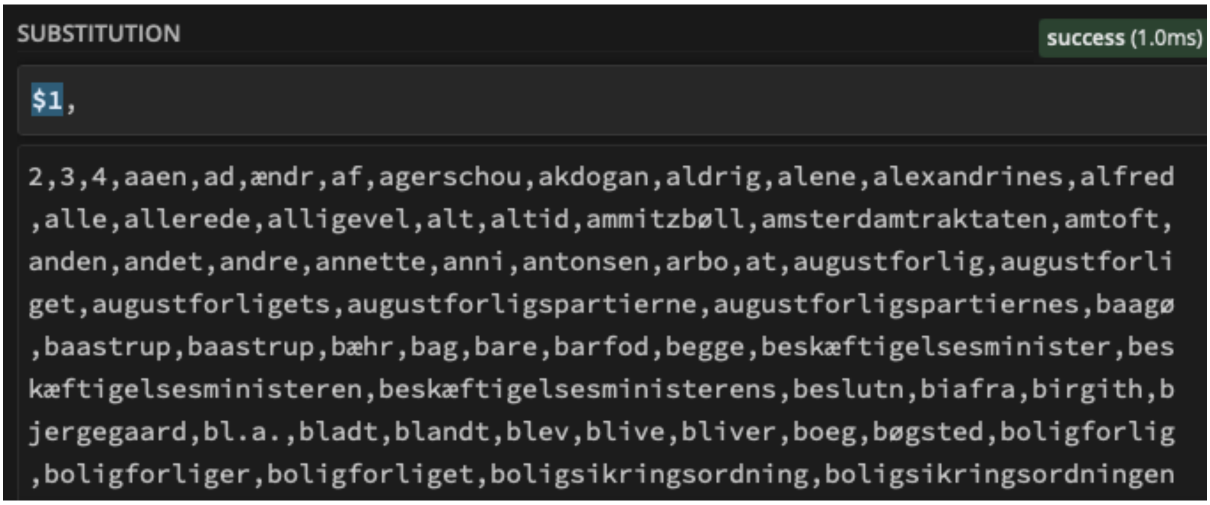
To convert the stopwordlist from Voyant into a stopwordlist for R, I used the regular expression:



```
/ (\S+)\n?
```

The expression matches any non-whitespace character which occur one or more times (\S+), and any new line characters which occur zero to one time are removed (\n?).

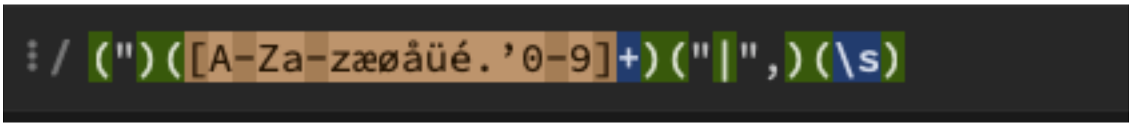
Under "substitution" I typed $1, to separate the words in the hyphen by the commas:



```
$1,
```

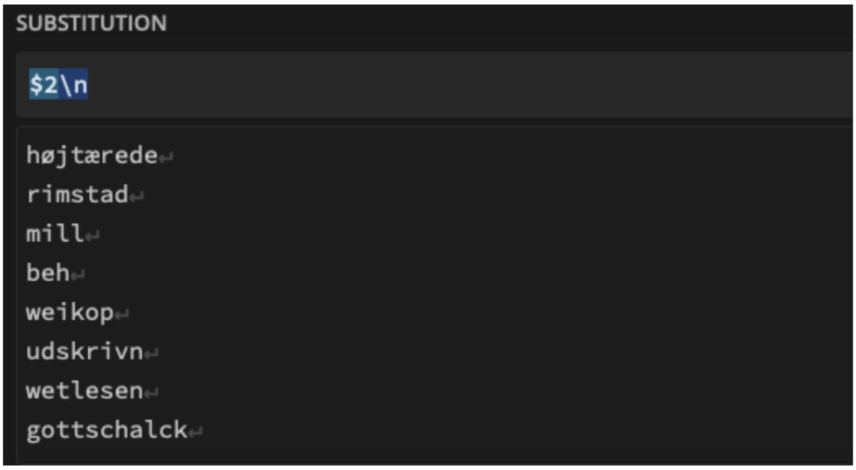
```
2,3,4,aaen,ad,ændr,af,agerschou,akdogan,aldrig,alene,alexandrines,alfred
,alle,allerede,alligevel,alt,altid,ammitzbøll,amsterdamtraktaten,amtoft,
anden,andet,andre,annette,anni,antonsen,arbo,at,augustforlig,augustforli
get,augustforligets,augustforligspartierne,augustforligspartiernes,baagø
,baastrup,baastrup,bæhr,bag,bare,barfod,begge,beskæftigelsesminister,bes
kæftigelsesministeren,beskæftigelsesministerens,beslutn,biafra,birgith,b
jergegaard,bl.a.,bladt,blandt,blev,blive,bliver,boeg,bøgsted,boligforlig
,boligforliger,boligforliget,boligsikringsordning,boligsikringsordningen
```

To convert the stopwordlist from R into a list for Voyant without punctuation, I used the regular expression:



```
/ (")([A-Za-zæøåüé.’0-9]+)("|",)(\s)
```

The second hyphen contains the actual words on the list, which must contain any capital or small letter and/or any digit one or more times. The words are separated from the punctuation by the regular expressions in the first and third hyphens. The expression in the fourth hyphen matches any space, tab or new line. I clicked on "substitution" and wrote $2\n to list the words inside the second parenthesis, and this removes the punctuation:



```
$2\n
```

```
højtærede
rimstad
mill
beh
weikop
udskrivn
wetlesen
gottschalck
```

Task 3: "What are the basic principles for using spreadsheets for good data organisation?"

When using spreadsheet for data organisation, there is a few things to be aware of. You shouldn't colour your cells as part of your data because the computer will not interpret a colour as data. If a piece of data has the value 0, you should write this in the cell and not just leave the cell empty, because this will be interpreted by the computer as missing data. If you actually have no data, then choose a name for this and be consistent. The name could be -999, NULL, NA or leaving the cell empty, but choose a name that your particular software understands. If a name of an object is misspelled, it will not be recognized as the correct object, and it will get its own category when statistics are made. Therefore, you must be consistent with variable names and not sometimes write "y" and other times "yes" if they are supposed to mean the same thing. Don't put several observations in one cell but give each observation its own column. Don't write additional notes outside the table but make the note into an observation with its own column. It is a bad idea to create multiple tables in one spreadsheet. Instead, give each case its own row in the same table to keep the data together. Make consistent column names, and don't name some with underscore and others without. Be aware if you create multiple tabs in one spreadsheet, because if you save the spreadsheet as a csv, the only tab which will be saved is the one you are currently working in.