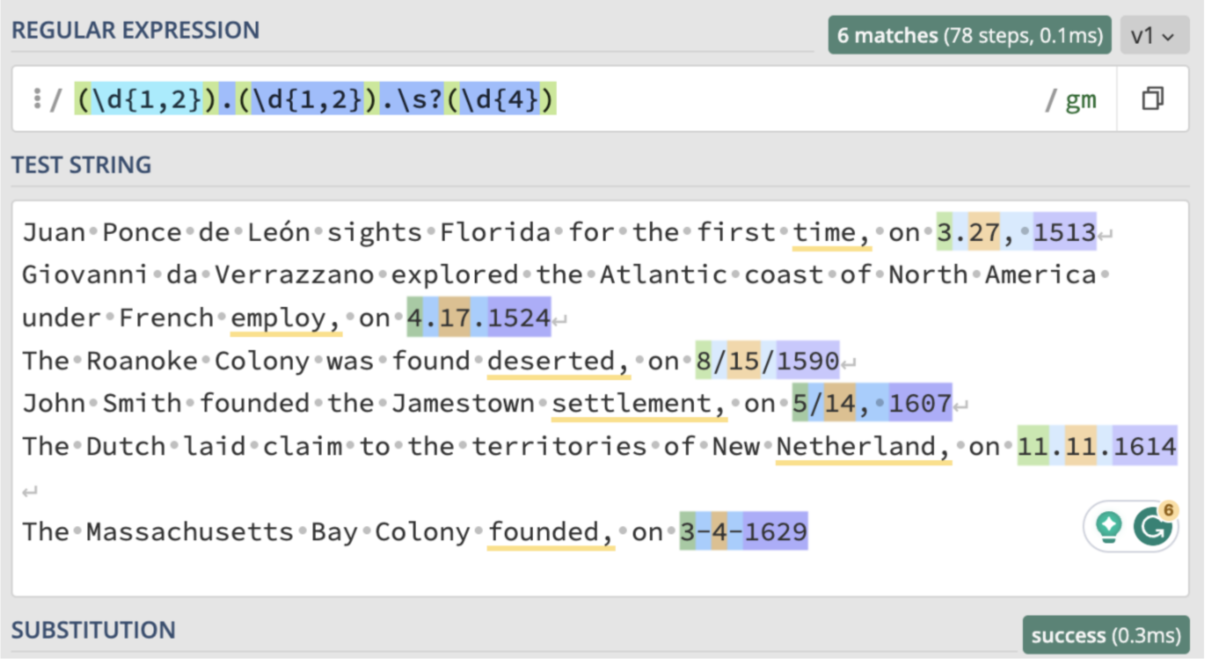
Clara Sydow
Digital Methods

Homework - regular expressions

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?**



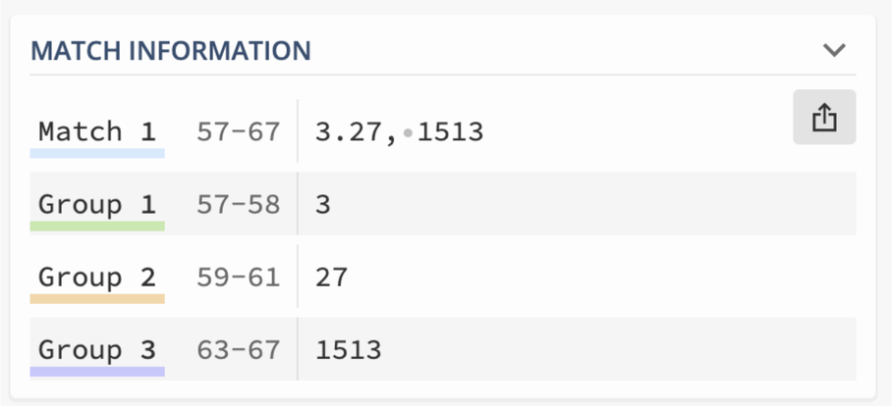
```
/ (\d{1,2}).(\d{1,2}).\s?(\d{4}) / gm
```

```
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
```

In the regular expression that I have used to extract the dates in the blurb and put them in the format: YYYY-MM-DD, I used the regex character: d{VALUE} to find the day, month and year. To find the day and month I used the value: 1,2 and to find the year I used the value: 4.

I used regular parentheses to isolate the day, month and year. between these, I used the character: . to match any character.
Before the year I used \s? to find any space, tab or newline that appeared either zero or one time.

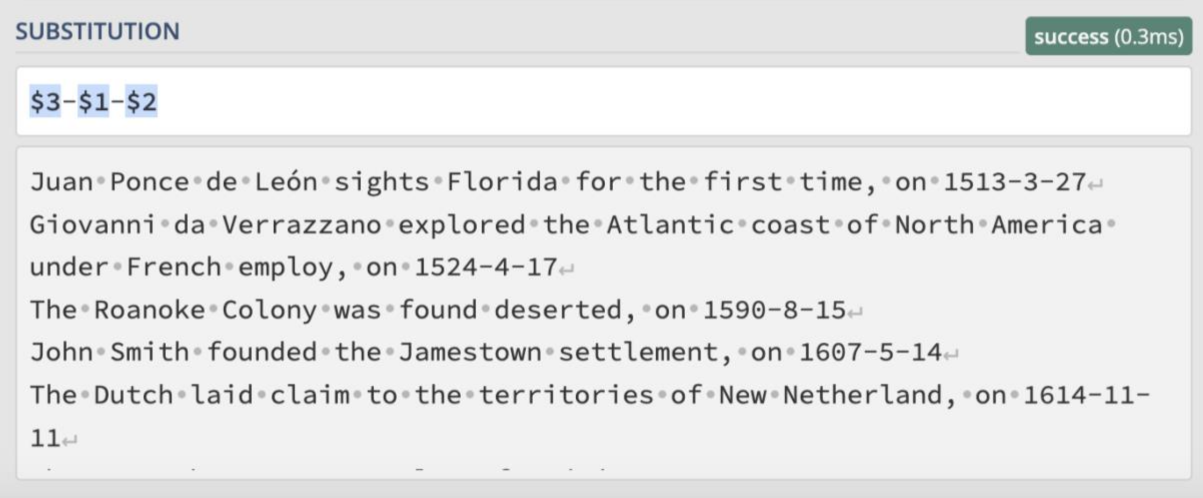


Match 1	57-67	3.27, 1513
Group 1	57-58	3
Group 2	59-61	27
Group 3	63-67	1513

To put them in the format: YYYY-MM-DD I used the function, substitution. Because I have isolated the day, year and month they have been categorized as group 1, group 2 and group 3.
Group 3: year

Clara Sydow
Digital Methods

Group 2: day
Group 1: month



```
$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
```

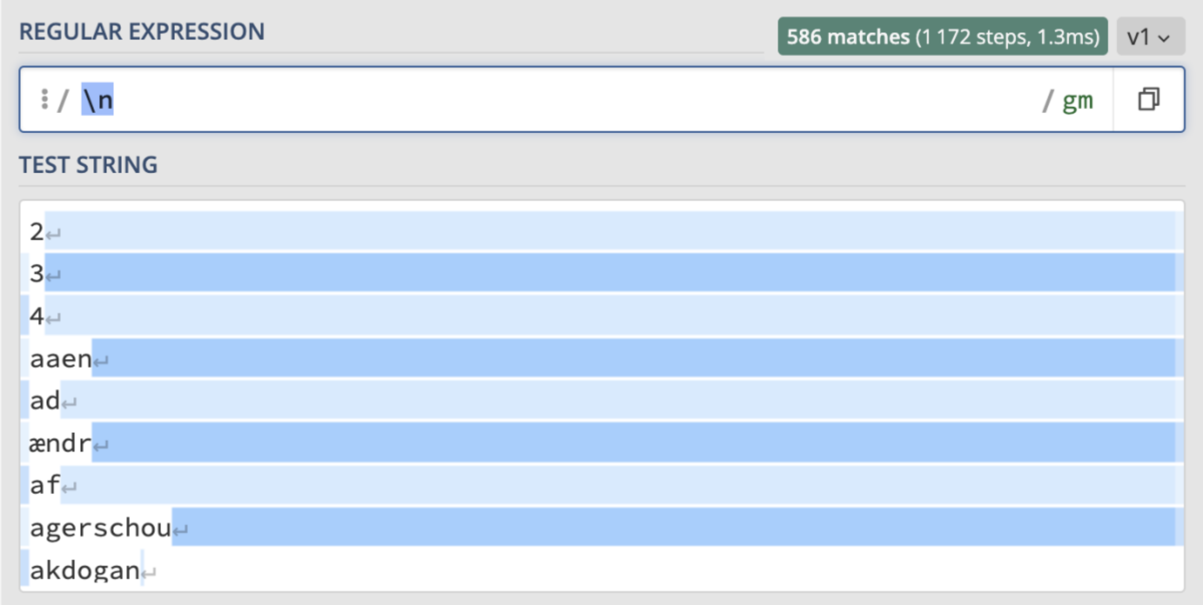
Therefore I can now put them in the order I want by writing: $3-$1-$2. The $ is used to capture the group that I am looking for.

Here is the link to my solution: https://regex101.com/r/diqHCZ/1

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)**

From Voyant to R:

The regular expression I used was: \n. This will capture all the space in the stopwordlist.

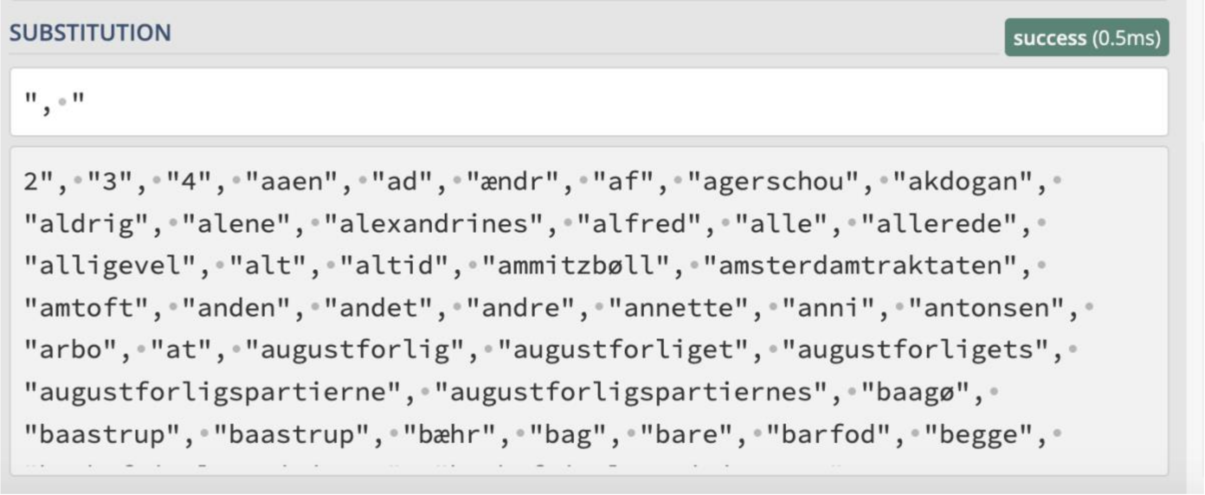


```
/ \n / gm
```

```
2
3
4
aaen
ad
ændr
af
agerschou
akdogan
```

Clara Sydow
Digital Methods

Afterwards, I used the function, substitution, to separate the “words” by commas. I used the following regular expression: “,•”
• = space



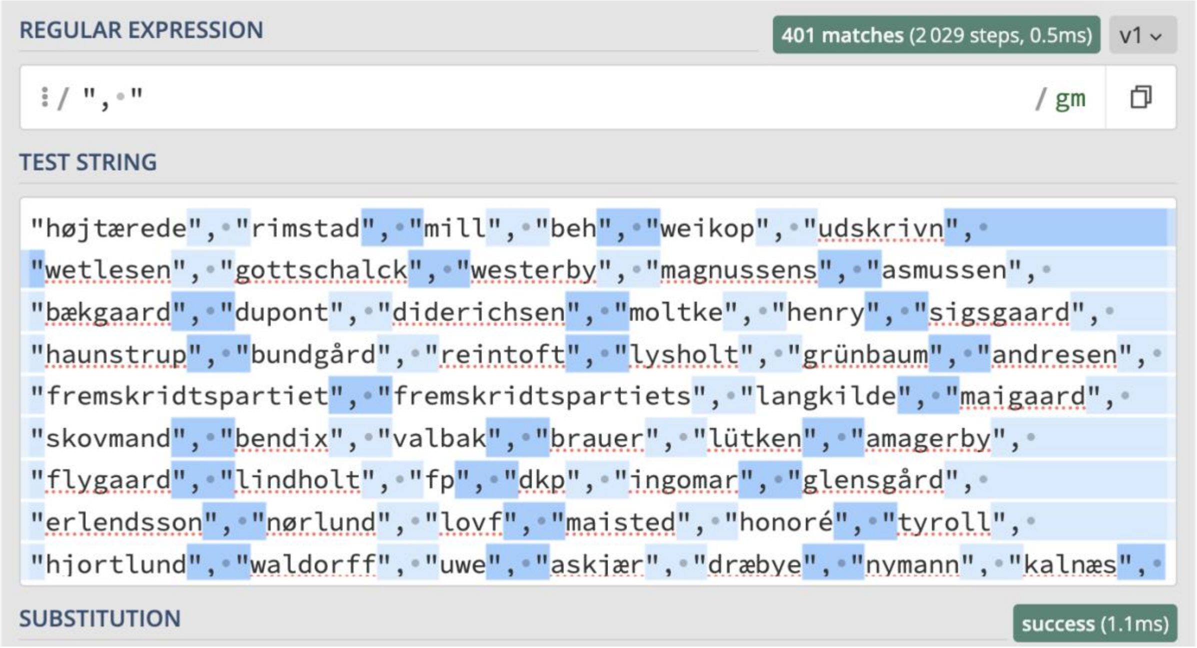
```
",•"
```

```
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",•"augustforliget",•"augustforligets",•
"augustforligspartierne",•"augustforligspartiernes",•"baagø",•
"baastrup",•"baastrup",•"bæhr",•"bag",•"bare",•"barfod",•"begge",•
```

Here is a link to my solution: https://regex101.com/r/tn0GTZ/1

From R to Voyant:

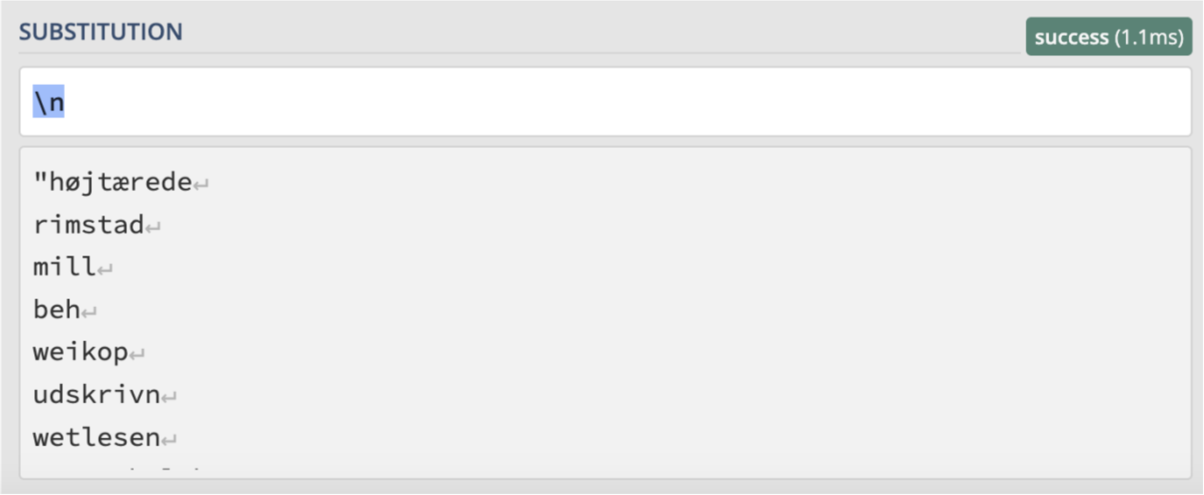
The regular expression I used was: “,•”. This will capture all the quotation marks, commas and space.



```
/ ",•" / gm
```

```
"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",•
```

Afterwards, I used the function, substitution, where I wrote the following regular expression: \n. This will substitute the “,•” with space and make it into a stopwordlist where the “words" appear on separate lines without interpunction.



SUBSTITUTION

success (1.1ms)

```
\n
```

```
"højtærede↵
rimstad↵
mill↵
beh↵
weikop↵
udskrivn↵
wetlesen↵
```

Here is the link to my solution: https://regex101.com/r/OgUFHq/1

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

The main principle for using spreadsheets for good data organisation is to be consistent. This applies to all the things you write and add to your spreadsheet. For example, when writing dates, be consistent in the way you do it: write YYYY-MM-DD. This is the global standard for how to write dates. Be attentive to the fact that Excel has some difficulties with dates. It can sometimes turn data into dates or turn dates into something else.

Do not leave cells empty in your spreadsheet. If you have missing data use a common code, for example, write "NA" or a hyphen. This makes it known to the reader, that you know that data is missing and therefore not left blank intentionally.
When talking about cells, the best layout for your spreadsheet is in the form of a rectangle, where the columns are the variables and the rows are the subjects. At the top of the rectangle, in the first row, the names of the variables should be written.

Consistency also applies to the names you give the categorical variables. Stick to one thing: either call it male/female or M/F, not both. To add to this, it is generally important to give good names to things in your spreadsheets. Do not use spaces in file names or in the names of the categorical variables. Use underscores or hyphens instead of space.

To sum it up, the keyword for the basic principles for using spreadsheets for good data organisation is consistency. And remember to make backups of your data!