1:W35: Regular expressions and spreadsheets

DESCRIPTION

Upload your answers/solutions to the problems below. Beware of making the submission legible and understandable to another reader:

1.1) Date matching and substitution

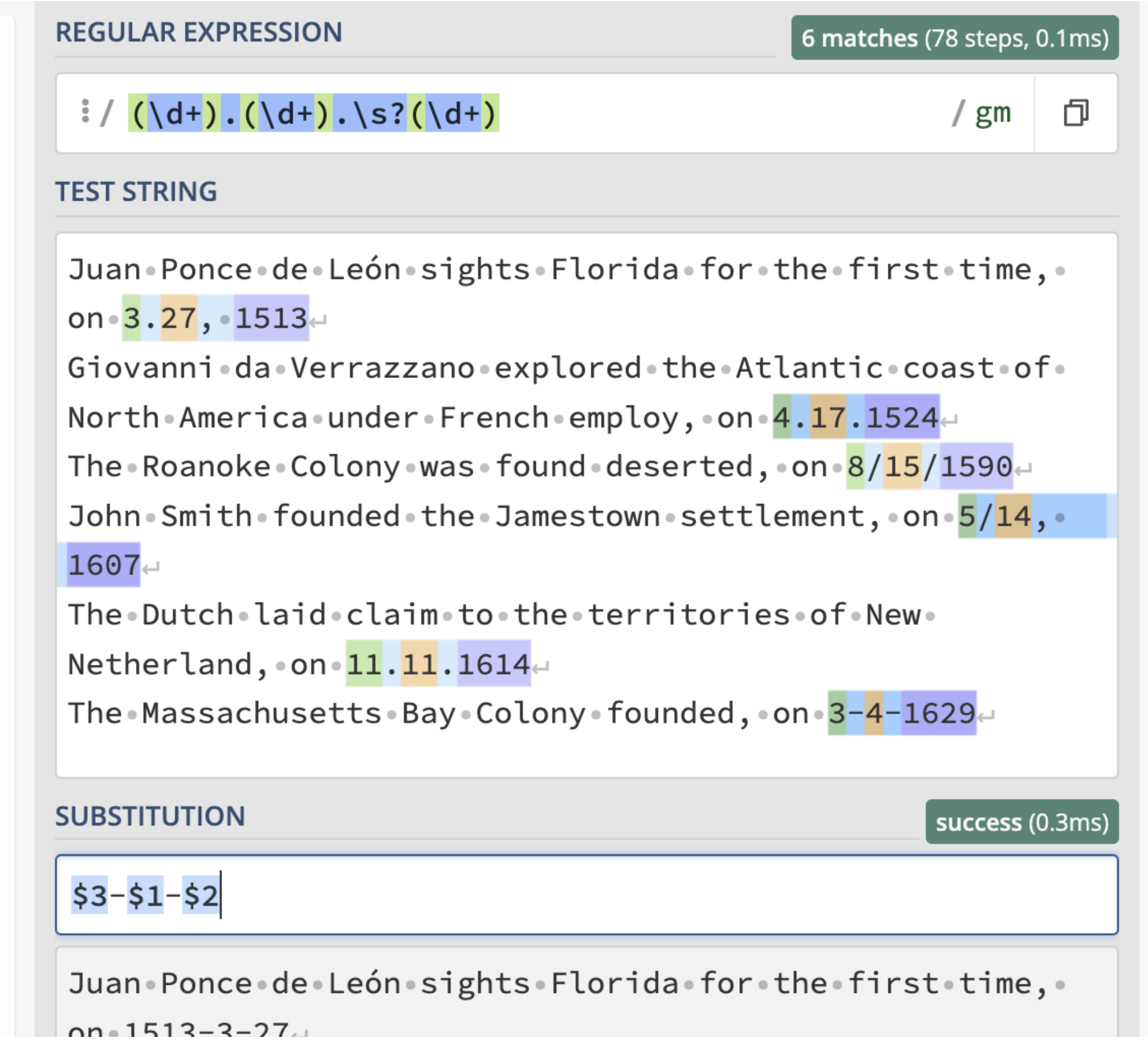
What regular expressions do you use to extract all the dates in this blurb: *http://bit.ly/regexexercise2* and to put them into the following format YYYY-MM-DD ?

To match all dates, I can use the following regular expressions: `\d+.\d+.*\d+` or `\d+.\d+.\s?\d`

To change format, I have to create variables. For this, we need to use round brackets/parenthesis (), indicating that everything within one set of parentheses should be regarded as one variable: `(\d+).(\d+).\s?(\d+)`

For instance, the bracket `(\d+)` means it is the first group and has a meaning of its own.

Having grouped the regular expression into 3 sets of parenthesis means that we have now isolated them as 3 different variables. I then click "substitution" in the regex101 interface and recombine them. When specifying the substitution, we refer to each group with a dollarsign and a number resembling the order of the 3 groups.



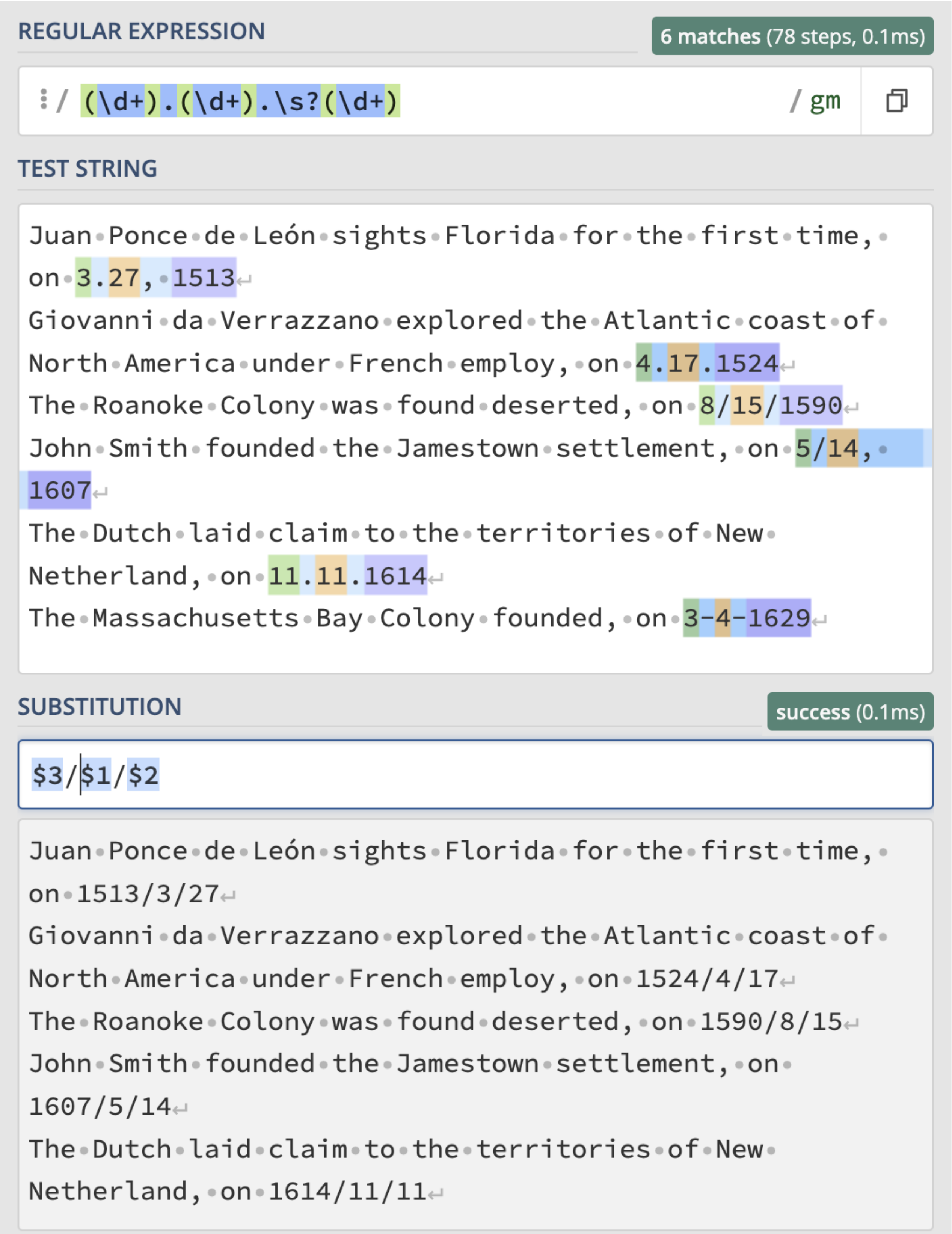
```
/ (\d+).(\d+).\s?(\d+) / gm
```

```
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
```

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

```
Juan Ponce de León sights Florida for the first time, on 1513-3-27
```

Or change the separator!



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

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

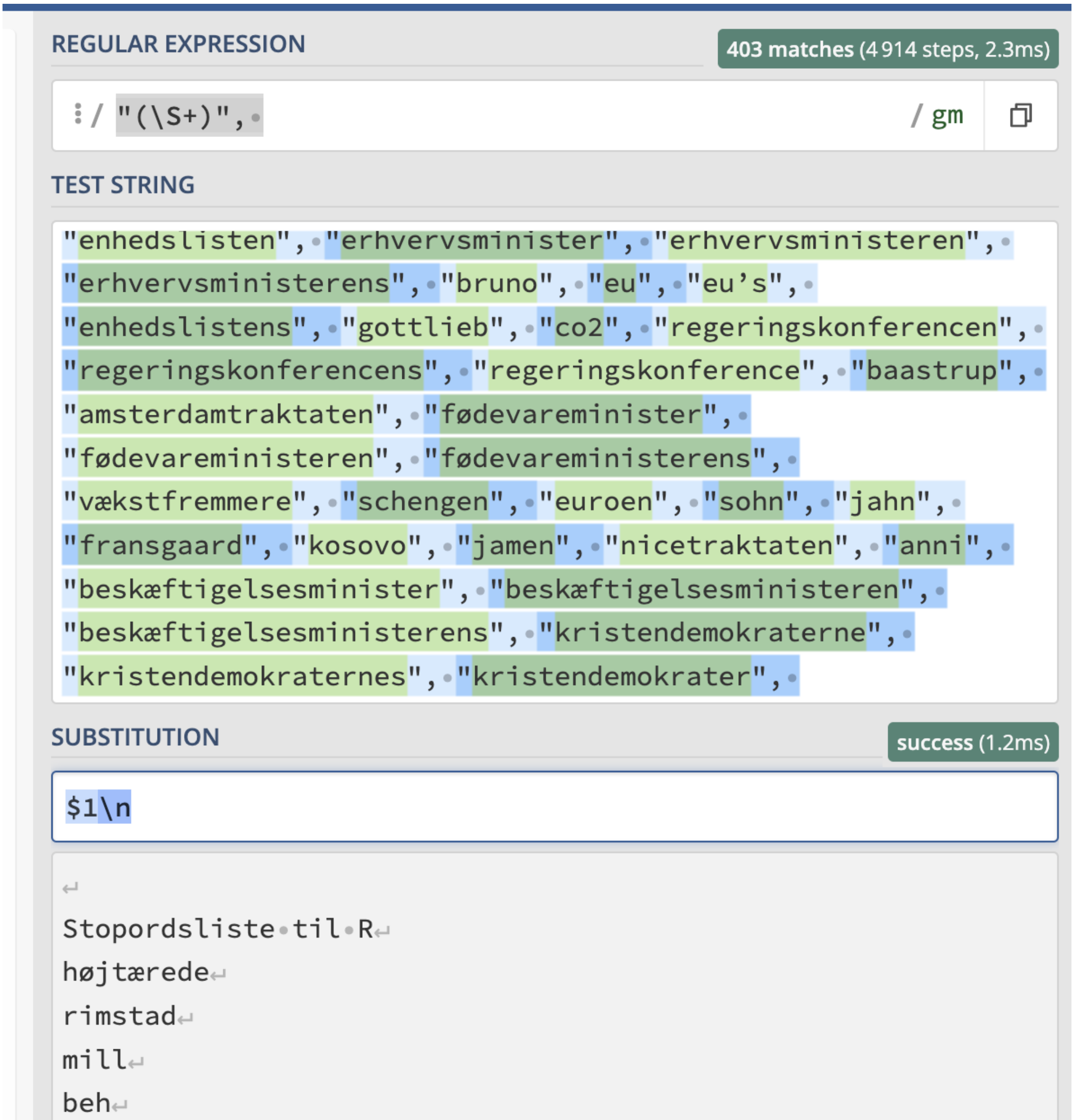
```
Juan Ponce de León sights Florida for the first time, 
on 1513/3/27
Giovanni da Verrazzano explored the Atlantic coast of 
North America under French employ, on 1524/4/17
The Roanoke Colony was found deserted, on 1590/8/15
John Smith founded the Jamestown settlement, on 
1607/5/14
The Dutch laid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

1.2) Stopwordlist for Voyant and R

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)

1.2.1 From text to column (From R to Voyant)

When doing text (R) to column (Voyant) we want to grab the group of `(", ")` and substitue with `\n`, i.e. a line break. When having done that, we copy the text string we have left and now just need to get rid of the first and last citation mark `"`. In the example below, we can also add an `?` in the end to catch the last word.



```
/ "(\S+)",· / gm
```

```
"enhedslisten",·"erhvervsminister",·"erhvervsministeren",·
"erhvervsministerens",·"bruno",·"eu",·"eu’s",·
"enhedslistens",·"gottlieb",·"co2",·"regeringskonferencen",·
"regeringskonferencens",·"regeringskonference",·"baastrup",·
"amsterdamtraktaten",·"fødevareminister",·
"fødevareministeren",·"fødevareministerens",·
"vækstfremmere",·"schengen",·"euroen",·"sohn",·"jahn",·
"fransgaard",·"kosovo",·"jamen",·"nicetraktaten",·"anni",·
"beskæftigelsesminister",·"beskæftigelsesministeren",·
"beskæftigelsesministerens",·"kristendemokraterne",·
"kristendemokraternes",·"kristendemokrater",·
```

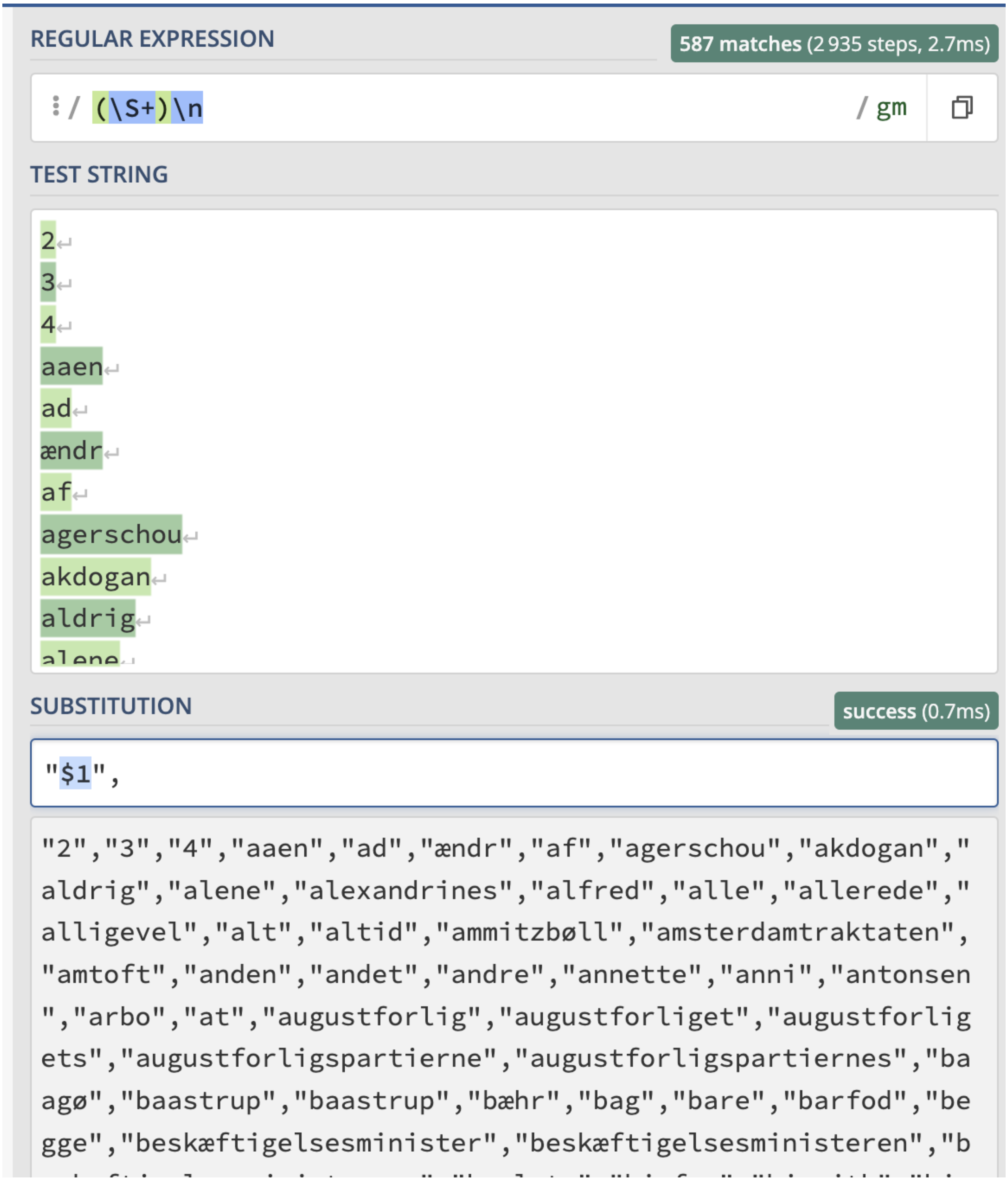
```
$1\n
```

```
↵
Stopordsliste·til·R↵
højtærede↵
rimstad↵
mill↵
beh↵
```

Another solution is `(\",")|(\"")` and then `\n` in the substitution field.

1.2.2 From column to text

Below regular expression can be used to take a column (Voyant) and turn into text (R). The space doesn't matter in R. We also have to remove the last comma. We can do that by matching a comma and substitute it with *nothing*.



REGULAR EXPRESSION

587 matches (2 935 steps, 2.7ms)

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

TEST STRING

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

SUBSTITUTION

success (0.7ms)

```
"$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","augustforliget","augustforlig
ets","augustforligspartierne","augustforligspartiernes","ba
agø","baastrup","baastrup","bæhr","bag","bare","barfod","be
gge","beskæftigelsesminister","beskæftigelsesministeren","b
```

1.3 Basic principles for using spreadsheets for good data organisation.

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

Data organisation in spreadsheets should follow principles as to minimise errors and alleviate later processes of data analysis. There are several simple, basic principles that ought to be followed.

First principle is to ***be consistent***. When making choice of formats, codes, etc. used for i.e., categorical variables or missing values, these codes should be consistent across the whole spreadsheet. This principle also applies to file naming and file layouts.

The second principle is to ***use good names*** and use underscores or hyphens instead of. Good names applies both to variables and to files. Variable names should make sense, and you should avoid spaces - and never include *'final'* in a file name.

Thirdly, ***the global ISO 8601 standard (YYYY-MM-DD) should be used for date formats***. Different countries have different ways of specifying dates, so following a global standard avoids confusion leading to errors.

Fourth principle is that ***no cells should be empty*** - in case of missing data, fill in a common code such as **NA** instead of leaving the cell blank. Also, each cell in a spreadsheet should contain only ***one piece of data****.

Additionally, the data layout should be a single rextangle with no more than one row for the variable names.

Lastly, one should create a ***data dictionary*** specifying variables, metadata, and a ReadMe file specifying content of the project and different files.

Lastly, det data should be tidy.

/ Source: Data Organization in Spreadsheets, The American Statistician, Broman & Woo 2017.

1.4 Optional: Dis Manibus

Challenge (OPTIONAL)! Can you find all the instances of 'Dis Manibus' invocation in the EDH inscriptions in *https://bit.ly/regexexercise5*? Beware of the six possible canonical versions of the Dis Manibus formula (see day 1 slides)!

Not full solution is the following RegEx expression `Di*s? M? ?\w+|`