1:W35: Regular expressions and spreadsheets

1. What regular expressions do you use to extract all the dates in this blurb: http://bit.ly/regexercise2 and to put them into the following format YYYY-MM-DD?

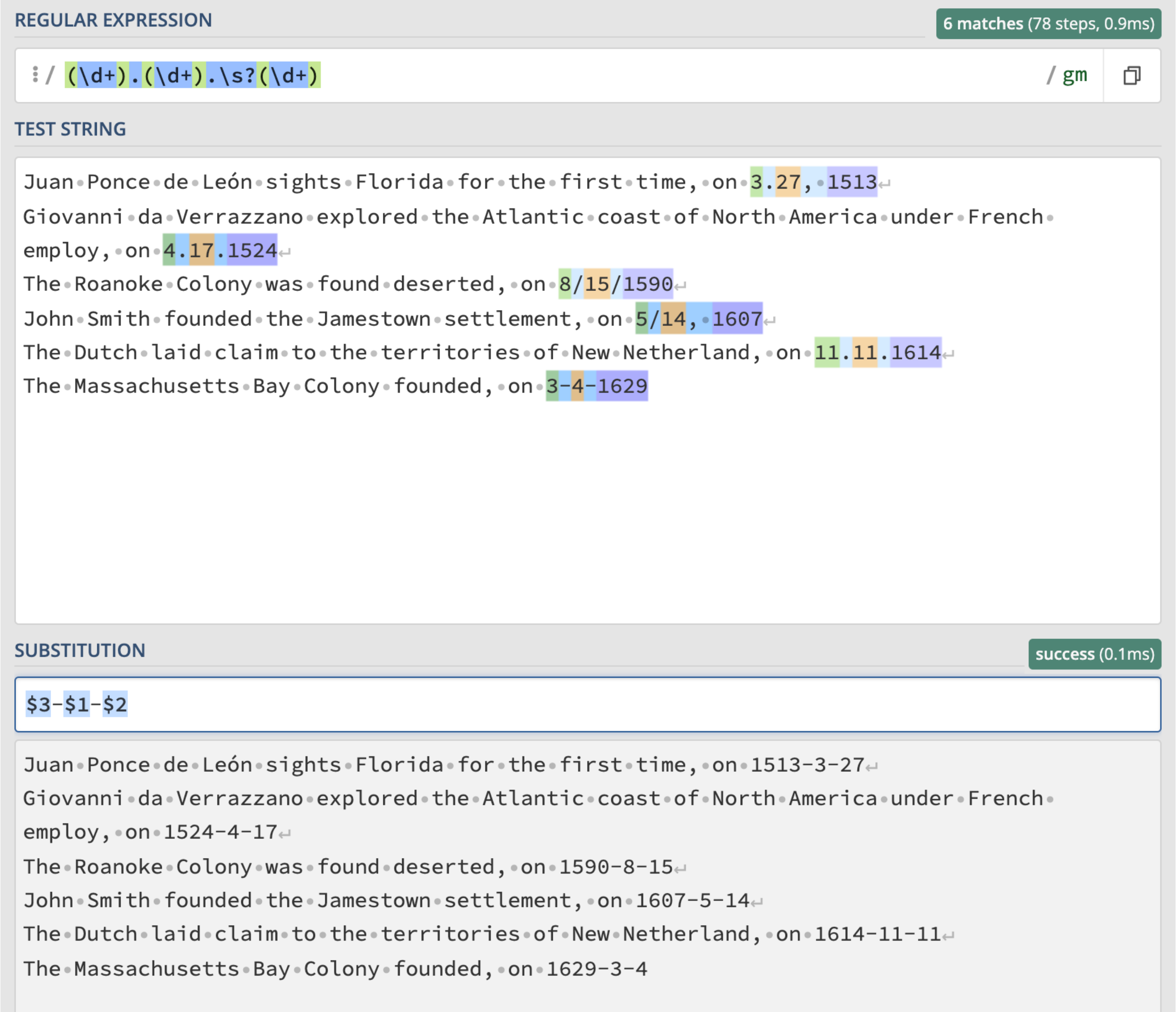
1.1 ANSWER

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

Substitution:

$3-$1-$2

Result:



REGULAR EXPRESSION

6 matches (78 steps, 0.9ms)

```
/ (\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 (0.1ms)

```
$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
```

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)

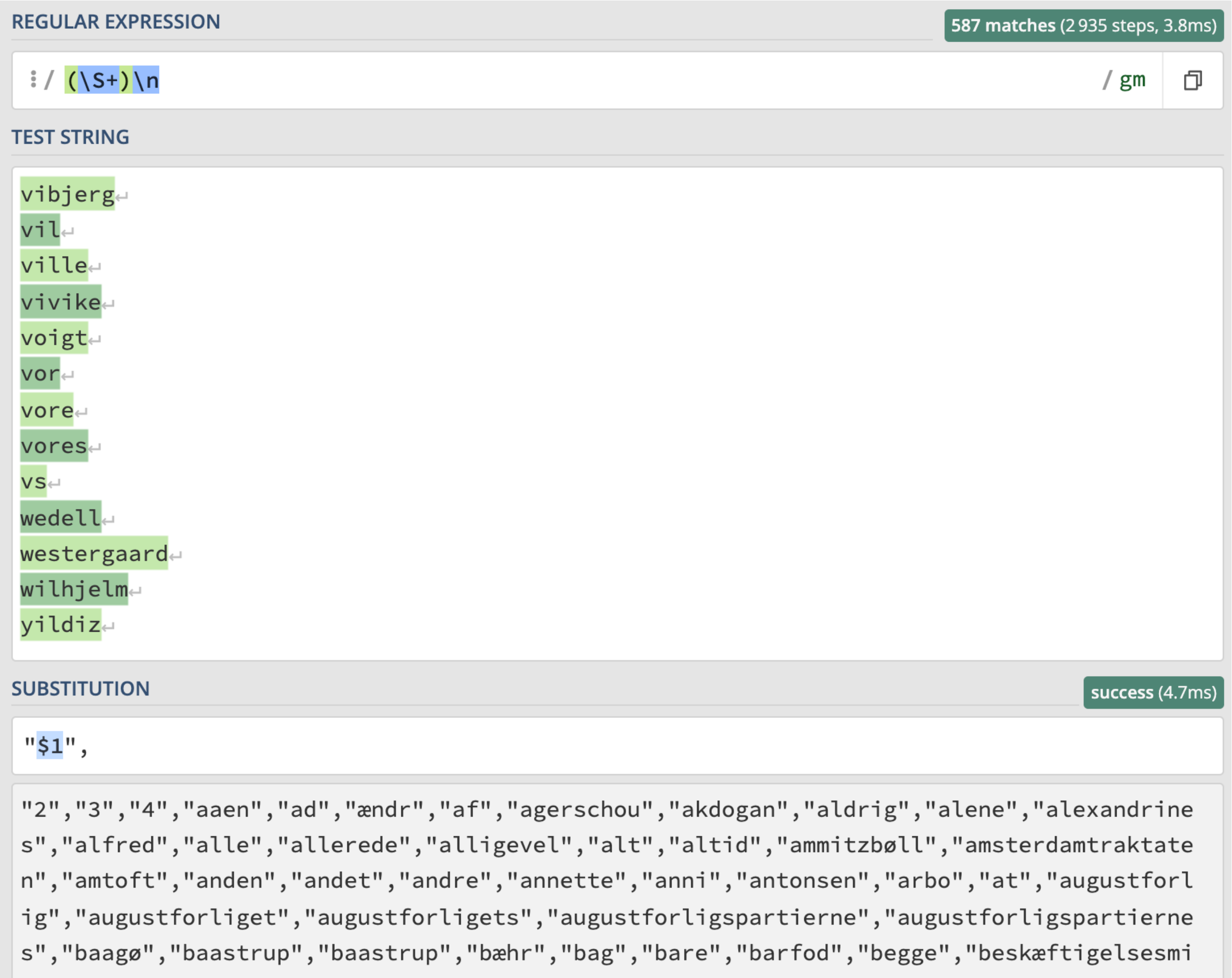
2.1 ANSWER – from Voyant list to R list

(\S+)\n

Substitution:

“$1”,

Result:



REGULAR EXPRESSION

587 matches (2 935 steps, 3.8ms)

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

TEST STRING

```
vibjerg
vil
ville
vivike
voigt
vor
vore
vores
vs
wedell
westergaard
wilhjelm
yildiz
```

SUBSTITUTION

success (4.7ms)

```
"$1",
```

```
"2","3","4","aaen","ad","ændr","af","agerschou","akdogan","aldrig","alene","alexandrine
s","alfred","alle","allerede","alligevel","alt","altid","ammitzbøll","amsterdamtraktate
n","amtoft","anden","andet","andre","annette","anni","antonsen","arbo","at","augustforl
ig","augustforliget","augustforligets","augustforligspartierne","augustforligspartierne
s","baagø","baastrup","baastrup","bæhr","bag","bare","barfod","begge","beskæftigelsesmi
```

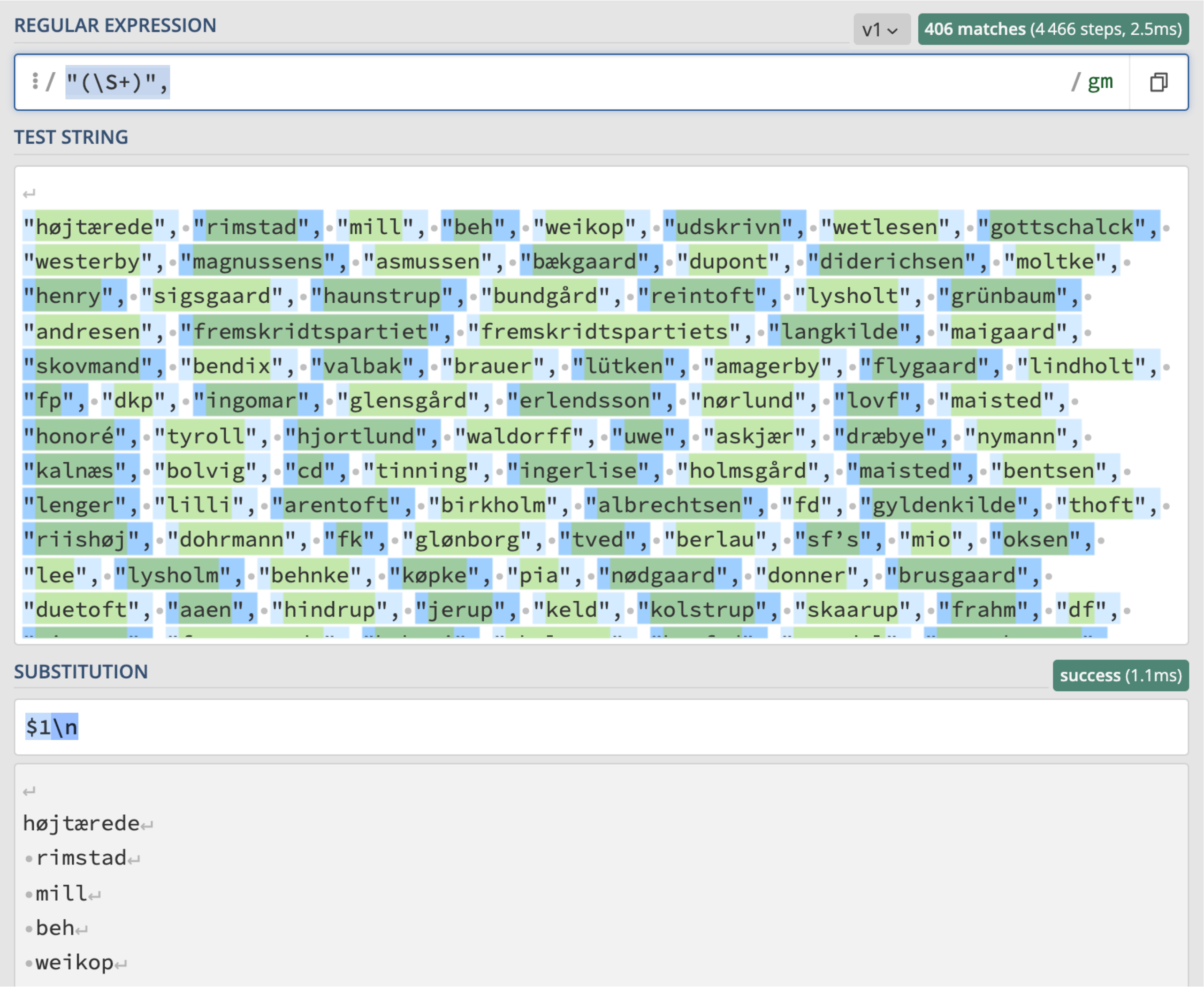
2.2 ANSWER – from R list to Voyant list

"(\S+)",

Substitution:

$1\n

Result:



```
/ "(\S+)", / 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", "bolvig", "cd", "tinning", "ingerlise", "holmsgård", "maisted", "bentsen", 
"lenger", "lilli", "arentoft", "birkholm", "albrechtsen", "fd", "gyldenkilde", "thoft", 
"riishøj", "dohrmann", "fk", "glønborg", "tved", "berlau", "sf's", "mio", "oksen", 
"lee", "lysholm", "behnke", "køpke", "pia", "nødgaard", "donner", "brusgaard", 
"duetoft", "aaen", "hindrup", "jerup", "keld", "kolstrup", "skaarup", "frahm", "df", 
```

```
$1\n
```

```
højtærede
 rimstad
 mill
 beh
 weikop
```

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

3.1 ANSWER

To ensure that one is using spreadsheets properly for good data organisation several things must be kept in mind. First, one should consider the software that will subsequently be fed the data: programs like R or Python cannot tell different tables in one sheet apart neither combine them to one, which makes it crucial to make your data well 'placed' and structured before entering another software in which the analysis will be done. This means that one sheet should consist of no more than one big data frame, in which all observations should be noted.

Consistency is key when notating each observation: all values in each column should be on the same format and no cell should be left empty. Additionally, only one value should appear in each cell and common notation should be used, for example dates should be noted like YYYY-MM-DD. One should never notate the measure of the observation inside the cell (like 0.8 sec) but instead use intuitive naming of the columns (like time_sec) and then leave the cell with the value only (like 0.8).

One should keep the data raw, and thus not use calculations or font or colour highlighting, since each step of the analysis should be transparent to others: each calculation step will not be saved in a typical sheet and the raw data will worst case be lost. The highlights will not be transferred into programs like R or Python anyways, and thus doesn't make sense to introduce anyway. Lastly, one should remember to make both backups and data validation and hereby ensure that data entry errors are avoided.