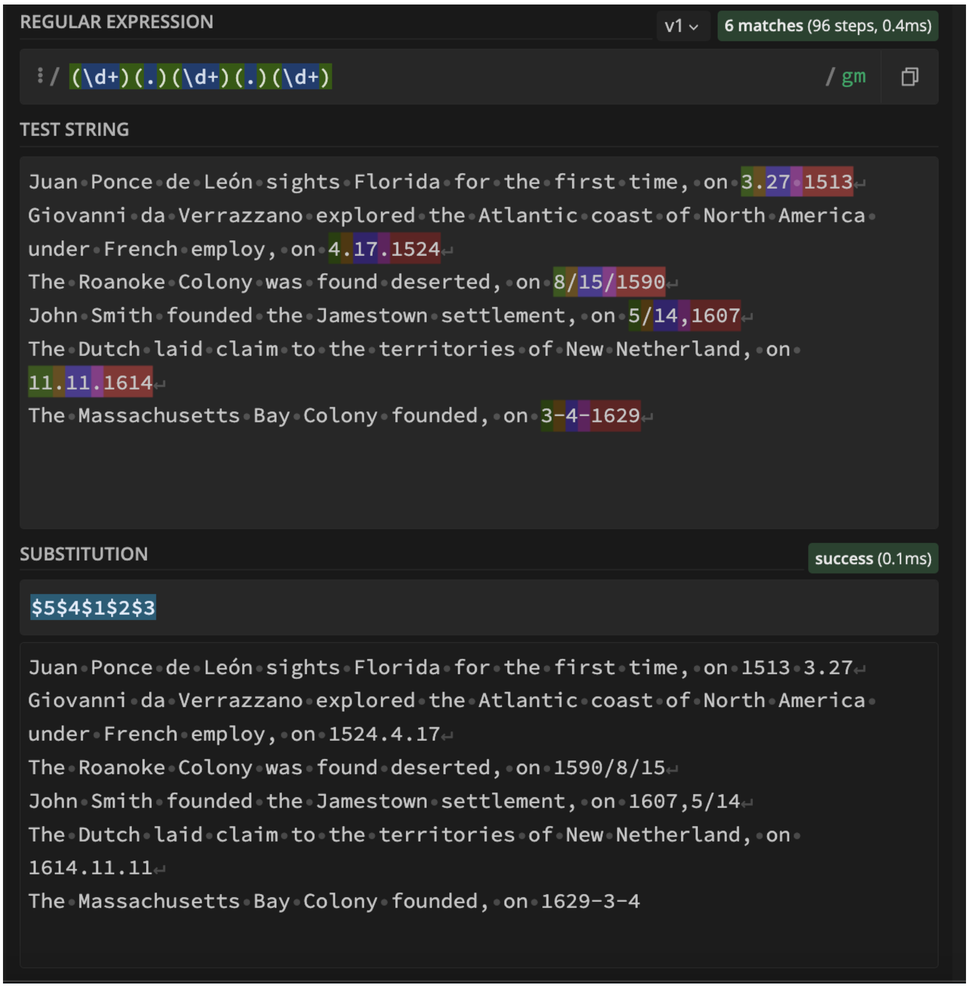
Cultural Data Science 2022
Rikke Uldbæk
202007501

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. **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 ?**
    a. https://regex101.com/r/TcWSuo/1



REGULAR EXPRESSION

6 matches (96 steps, 0.4ms)

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

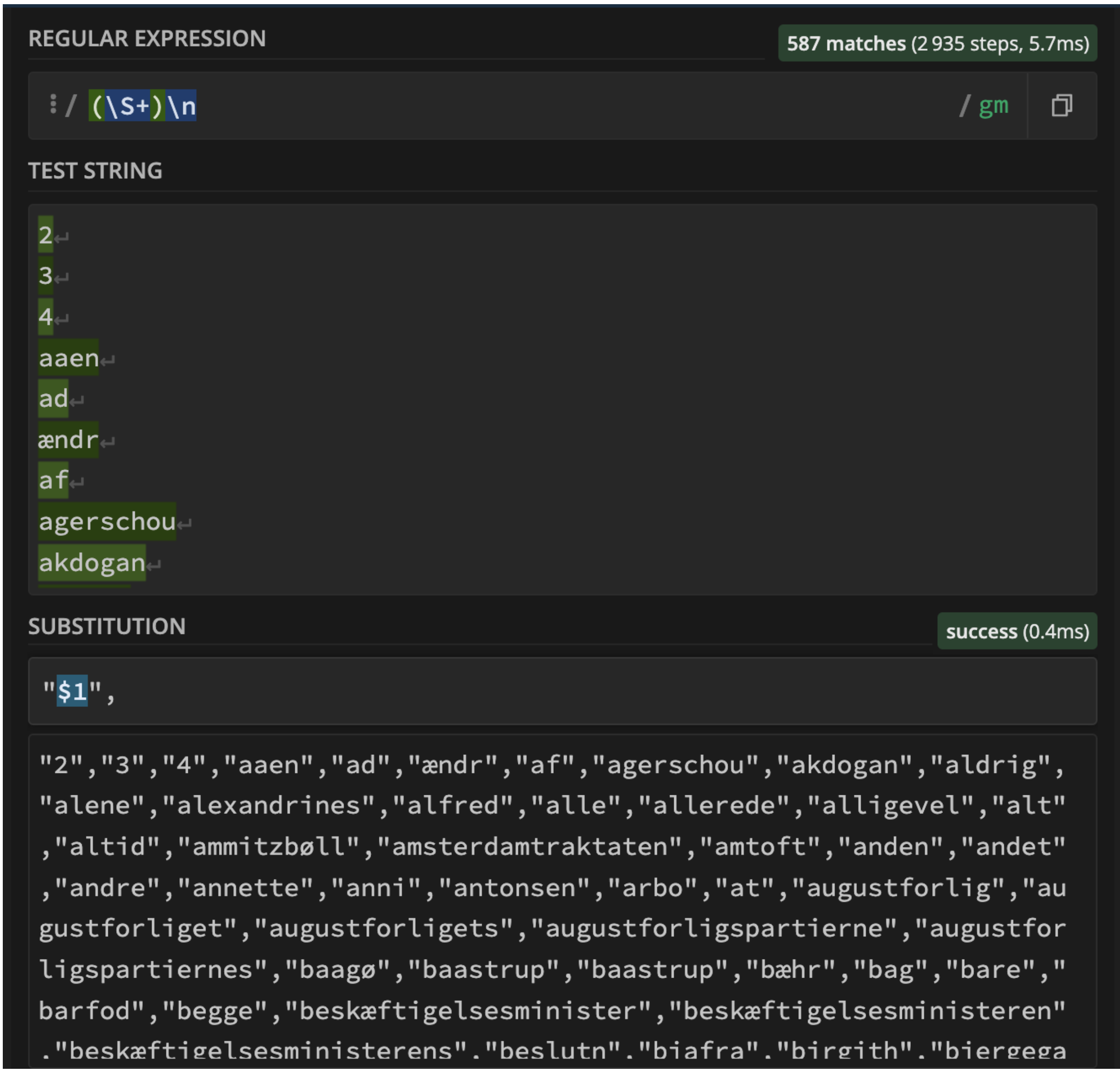
```
$5$4$1$2$3
```

```
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
```

    b.

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).**

Cultural Data Science 2022
Rikke Uldbæk
202007501

a. 

REGULAR EXPRESSION 587 matches (2 935 steps, 5.7ms)

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

TEST STRING

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

SUBSTITUTION success (0.4ms)

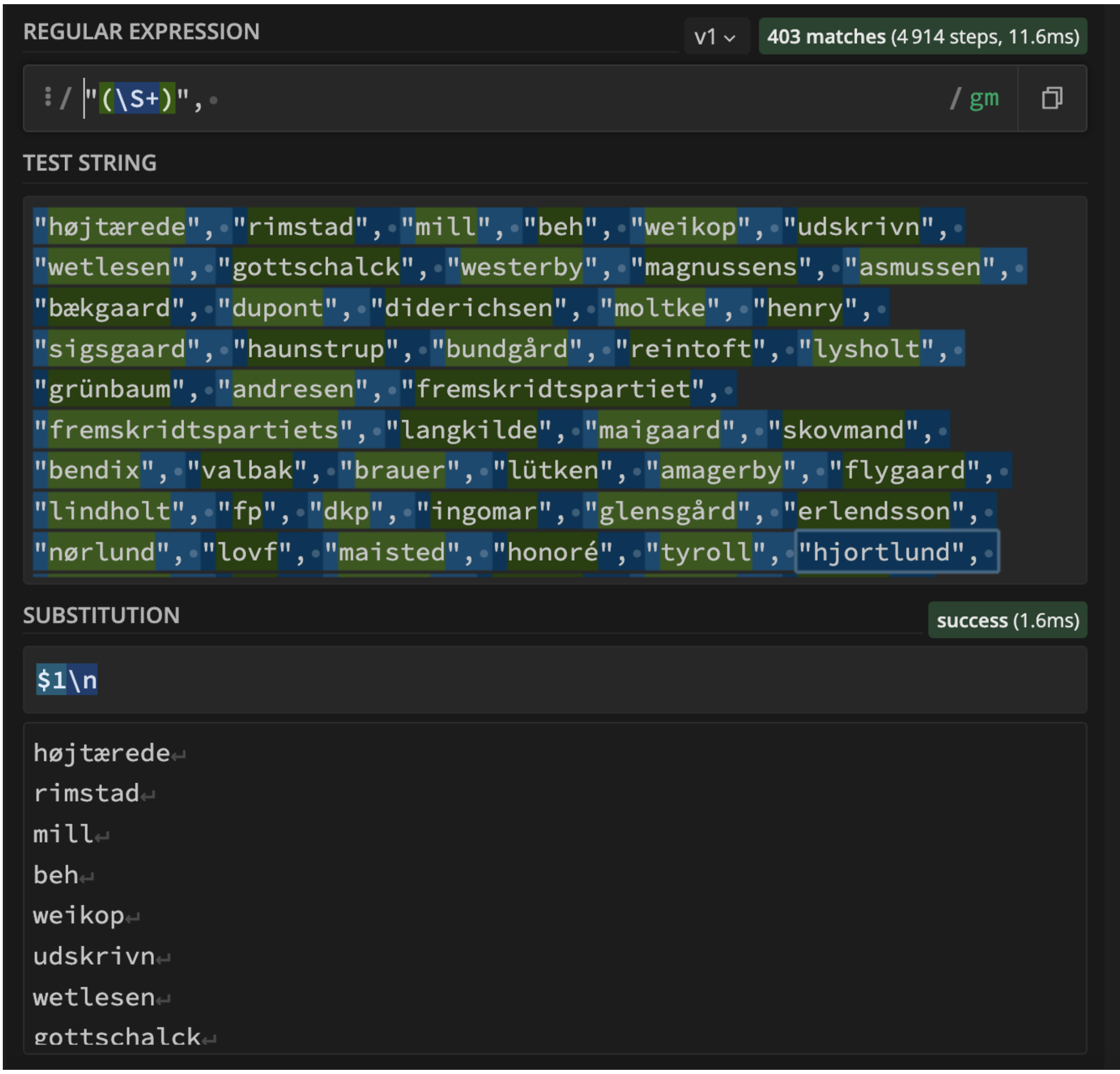
```
"$1",
```

```
"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","au
gustforliget","augustforligets","augustforligspartierne","augustfor
ligspartiernes","baagø","baastrup","baastrup","bæhr","bag","bare","
barfod","begge","beskæftigelsesminister","beskæftigelsesministeren"
,"beskæftigelsesministerens","beslutn","biafra","birgith","bjergega
```

b. https://regex101.com/r/HmUFTX/1

3. **Then take the stopwordlist from R http://bit.ly/regexexercise4 and convert it into a Voyant list (words on separate line without interpunction)**
    a. /library/SBeco4

Cultural Data Science 2022
Rikke Uldbæk
202007501



REGULAR EXPRESSION 403 matches (4 914 steps, 11.6ms)

```
/ "(\S+)", / gm
```

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", 
```

SUBSTITUTION success (1.6ms)

```
$1\n
```

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

b.

4. **In 250 words, answer the following question: "What are the basic principles for using spreadsheets for good data organisation?"**
    a. In order to obtain a neat and fine looking spreadsheet, a few pitfalls must be avoided. The following pitfalls need to be avoided; For instance it is important to name the columns something simple, short and indicative, using underscore instead of spaces if having more than one word. However, data within the rows containing multiple words, e.g. text strings, does not necessarily have to be separated by underscore, here normal spaces are fine. If any missing values are present in the spreadsheet, it's important to replace them with a consistent replacement, and here NA or NULL are the best replacements. Using anything other than NA and NULL, e.g., -999 or 0 is not a great idea, as these suddenly signify a meaningful value instead of a missing value. This can potentially harm a statistical analysis and create

falsely skewed results. Likewise, should each row just contain one data point and not multiple data points, of course dependent on the data type. To overcome this, it’s often better just to make dummy variables and extra columns. Misspellings and consistent naming of observations should be avoided to maintain reasonable spreadsheets.Thus, good practice when organizing and constructing spreadsheets is to be consistent, precise, concise, and unambiguous.

5. **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)!**