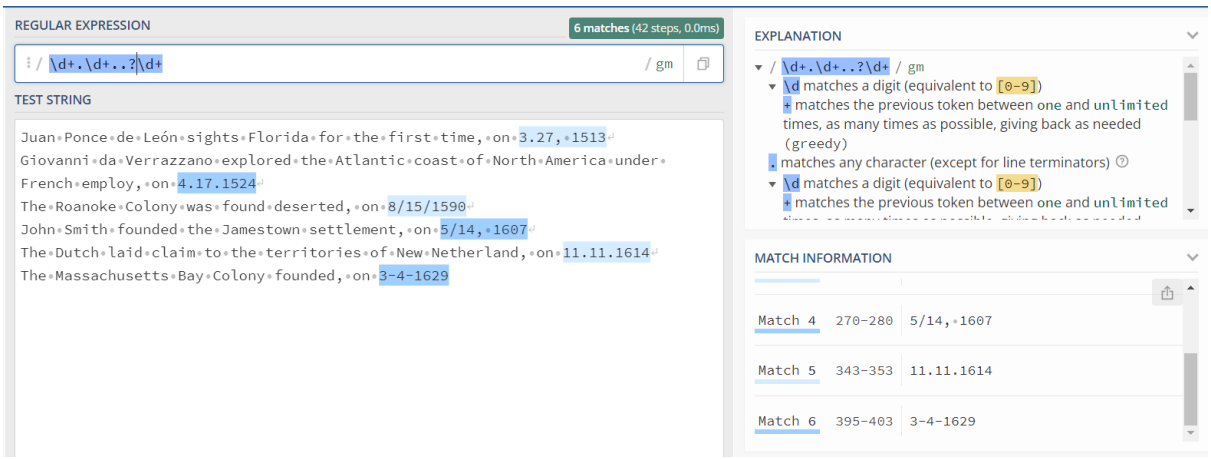
W44_Assignment_Mathias_Thomasen_Madsen

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?**

Since we in this assignment are to find the specific dates involving numbers, I started out using \d hence getting all the results for all the single digit numbers. Adding a + so it became \d+ gave me all the results for alle the numbers. The challenge was now to combine the single numbers into dates that followed the requirements given in the assignment (YYYY-MM-DD). Knowing that there would have to be three combined numbers to create the dates, the next step was writing \d+\d+\d+ which only gave the results of the years, but not the months and dates. Looking at the text set I saw that there were characters in between the numbers, hence I knew I in my regular expression would have to add something so that it would count these characters with it. I used the . to fill the space of these characters, and by trying I reached the regular expression of \d+\d+..?\d+., that answers this assignment:

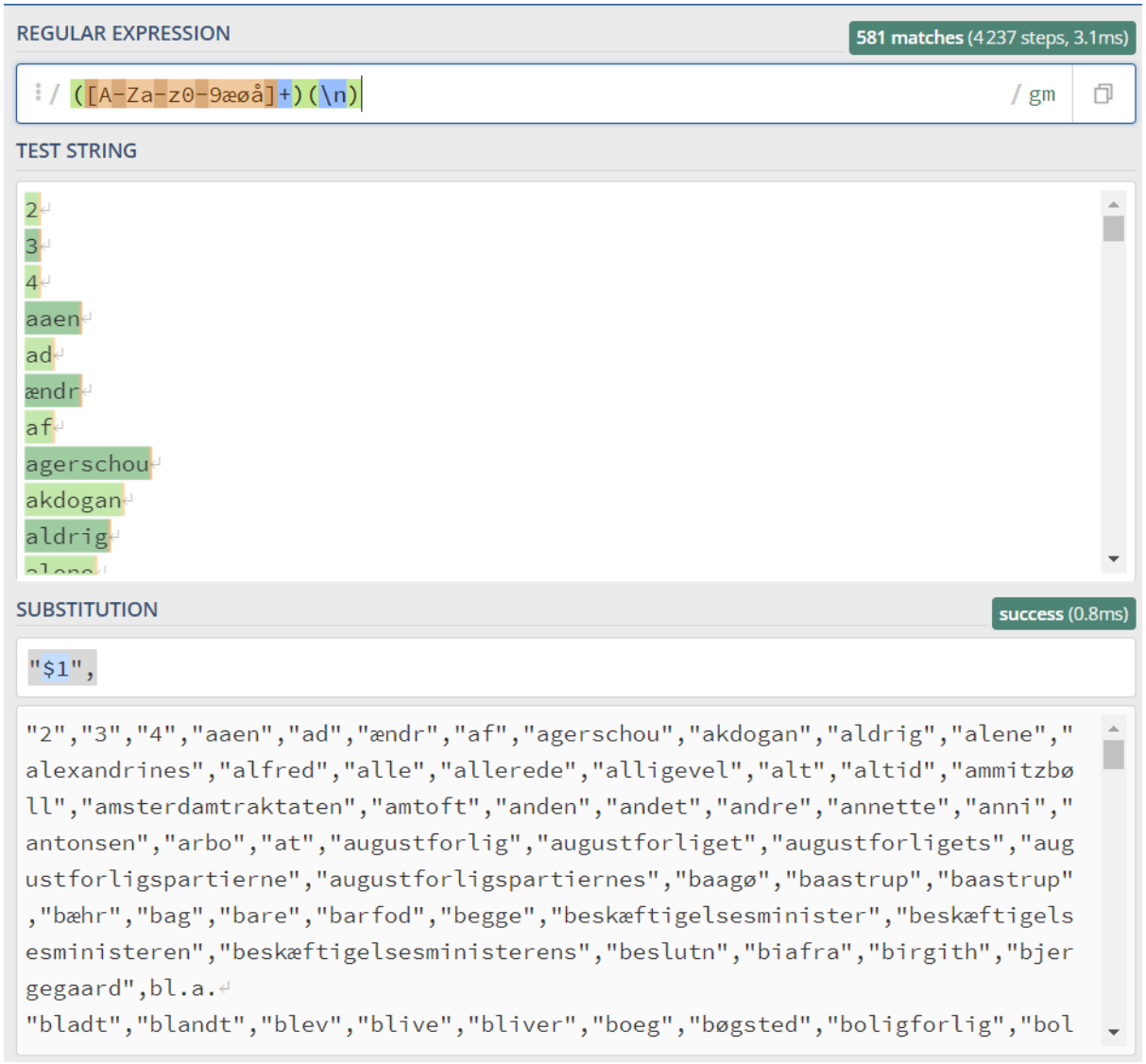


2. **Write a regular expression to convert the stopwordlist (list of most frequent Danish words) from Voyant in http://bit.ly/regexexercise3 into a neat stopword list for R (which comprises "words" separated by commas, such as http://bit.ly/regexexercise4). Then take the stopwordlist from R http://bit.ly/regexexercise4 and convert it into a Voyant list (words on separate line without interpunction)**

This is a two-part assignment. The first part is to convert the http://bit.ly/regexexercise3 stopwordlist into a stopwordlist for R. The second part is to convert the new R stopwordlist from the first part into a stopwordlist for Voyant.

First part:

I started of copying the dataset from http://bit.ly/regexexercise3 into regex101.com. The first thing I did in the regular expression was to match all the words, for which I entered \w+. Here I found, that \w does not include the Danish letters of æøå. Hence, I changed the regular expression to [A-Za-z0-9æøå]+ so that it would include all the words. The next step was to open the substitution and add the commas between all the words, with the expression "$1",. But when I opened the substitution and entered "$1", I found that all the words would be on separate lines. By looking at the dataset, I realized, that all the stopwords were on new lines. I searched on the internet and found, that by including the expression \n (\n is the expression for new line) in my regular expression, I got the stopwordlist for R that was required in the assignment:

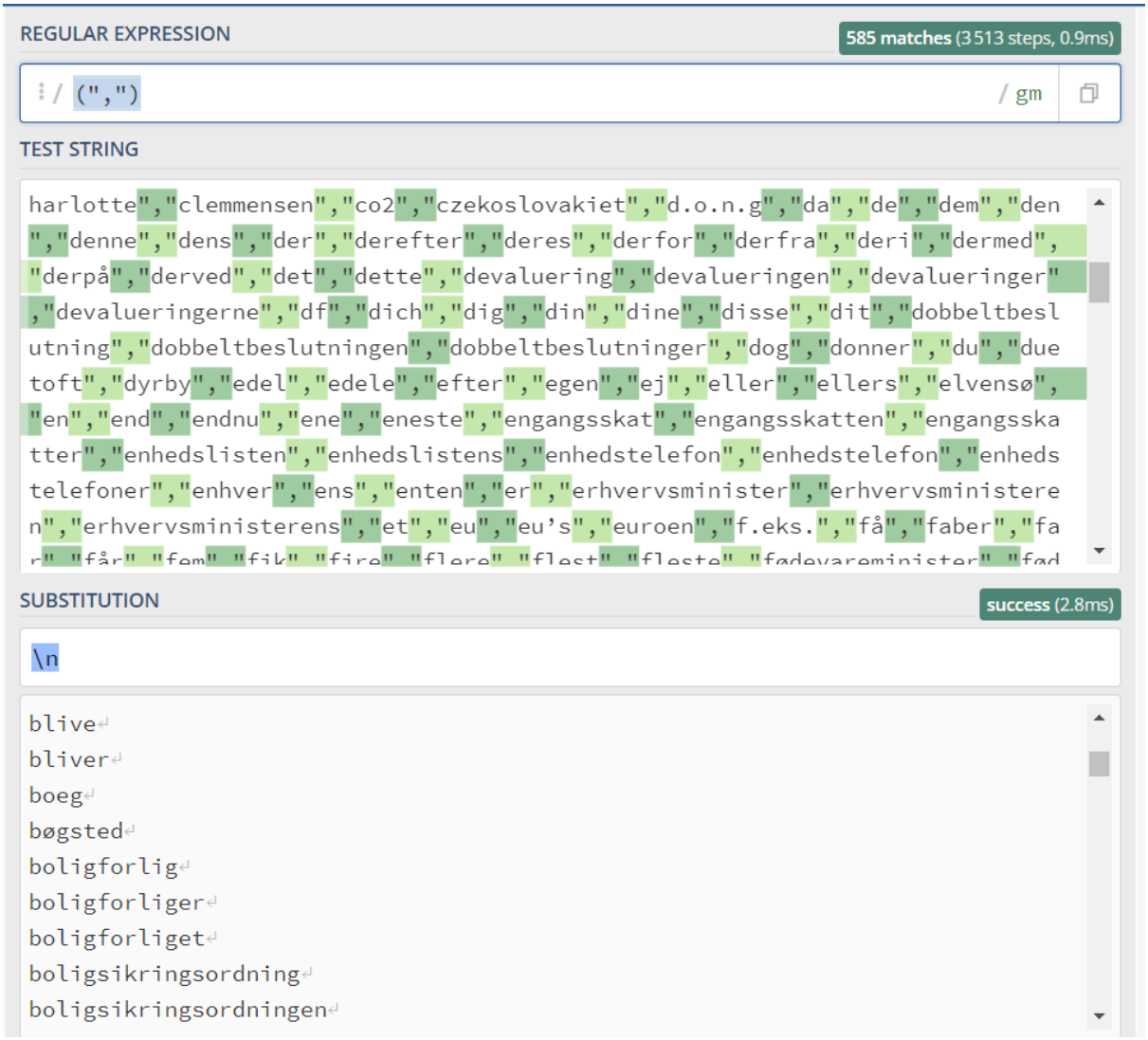


Second Part:

Now that I made the stopwordlist for R in the first part, I copied it into a new regex101.com. The first thing is to remove all these symbols from the text:

, . “

After trial and error, I found that writing the regular expression (",") would match the symbols. The next step was then to open the substitution and make it so, that every word would stand on its own line. In the first part of the assignment, I found that \n meant new line. By typing \n in the substitution, it resulted in a stopwordlist for Voyant, which was required in the assignment:



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

The main purpose of a spreadsheet is to later use it when making an analysis. When entering and creating the spreadsheet you should always have in mind, how you can make it easier for

the person utilizing the spreadsheet in an analysis later. Even though modern spreadsheets have multiple applications, you should only use it to store raw data, when doing so. You can create a new spreadsheet file where you may utilize the other tools to for example analyze the data. The main rule though, is to keep the spreadsheet with raw data separate from the analysis. The reason for this is, when utilizing other tools on your raw data, you may do something wrong and hence corrupting the data, so that it is all gone beyond saving. There are some ways to organize the spreadsheet or spreadsheets, so that they are easier to use for an analysis later. The main advice is: keep the spreadsheet clean, visually manageable and stay consistent.

The reason you want to stay consistent is, that when you use the same data layout, the different spreadsheets that you make will be instantly compatible. Also, it'll be easier to manage and organize the data when the files are stored in a consistent organized fashion. Other principles are to never have empty cells (write NA instead), and never write multiple things in one cell, because, at that point it's better to have more columns. Lastly, instead of using space between words when naming, use underscore. It's easier to work with.