W44: 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**](http://bit.ly/regexexercise2)**and to put them into the following format YYYY-MM-DD ?**

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

The regular expression used to extract all the dates in the text is: \d{1,2}.\d{1,2}..?\d{4}

\d matches any single digit followed by {1,2} which defines the digit as any number between one and nine digits in length. This signifies the month. This is followed by at punctuation mark which matches any character. This is then followed by a \d{1,2} which signifies the day again followed by a punctuation mark and a question mark which matches the preceding character zero or one time. The year is defined by a digit \d in the length {4}.

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

In order to convert the dates into the format YYYY-MM-DD, I will need to use to the substitution tool. The substitution tool allows me to substitute the data I have pulled from the text, with anything I wish to type instead. I can thus insert the dates in the requested format.

I order to substitute the dates with the required format, I firstly need to group the regular expression: (\d{1,2}).(\d{1,2})..?(\d{4})

I now have three groups of digits, $1 (MM), $2 (DD) and $3 (YYYY). I then use the substitution tool to insert the dates in the format YYYY($3)-MM($1)-DD($2):

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

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

Conversion from Voyant stopwordlist into R stopwordlist:

In order to convert the stopwordlist from Voyant to R, I will need to change the way in which the list is set up. In Voyant the words are separated by new lines, whereas in R, the words are separated by quotation marks as well as commas (“,”).

I need to identify the regular expression for new line, which is \n. I then need to use the substitution tool and substitute the new lines with quotation marks and commas. To get the right format I then simply need to add a quotation mark before the first word and after the last word.

**Et billede, der indeholder tekst

Automatisk genereret beskrivelse**

Conversion from R stopwordlist into Voyant stopwordlist:

In order to convert the stopwordlist from R into Voyant I will use the same strategy as above.

I need to identify the regular expression for quotation marks and commas, which is literally ",.". I then need to use the substitution tool and substitute the quotation marks and commas with new lines. To get the right format I then simply need to delete a quotation mark before the first word and after the last word.

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

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

Spreadsheets are a widely used tool for dealing with big data but can be associated with uncertainty. This requires the users to consider the basic principles for using spreadsheets to reduce errors. The first rule for using spreadsheets is consistency, as this will ensure a precise calculation. It is important to choose good and exact names to keep a clear overview when dealing with big data. When entering dates into spreadsheets, the recommended format is YYYY-MM-DD. Furthermore, it is important to leave no empty cells because they may cause confusion as to whether something is missing. Instead, use a consistent combination such as ”*NA*” to make it clear that the cell is meant to be empty. It is also important that the cells only contain one piece of information. Regarding layout, make a rectangle with rows equaling to subjects and columns equaling to variables. This is a commonly used layout as it offers a clear overview and if used consistently, makes it easy to transfer data. It might also be useful to create a data dictionary which explains the variables and eliminates doubt. Although font and highlighting colors may be productive, it is recommended to make an additional column which indicates deviations. When dealing with raw data, it is important to save a document containing untouched data as well as making regular backups in multiple locations. It is also a good idea to keep a copy of the data files in a plain text format. Lastly, it is recommended to use tools for avoiding data entry errors such as the tool offered in Excel, “data validation.”[[1]](#footnote-1)

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

1. Broman, Karl W., and Kara H. Woo. 2017. “Data Organization in Spreadsheets.” The American Statistician 72 (1): 2–10. https://doi.org/10.1080/00031305.2017.1375989. [↑](#footnote-ref-1)