HW1-W35 Regular expressions and spreadsheets

Author: Mina Almasi
Date: 1st of September, 2022

(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 ?

The REGEX is saved and can be accessed through the link HTTPS://REGEX101.COM/R/VSIKCS/1. The explanation is also below:

We search for all dates with the following:
`\d+.\d+.\s?\d+`

To put them into the format YYYY-MM-DD, we first use the parenthesis () to mark the different formats:
`(\d+).(\d+).\s?(\d+)`

To substitute (in regex101.com), we click the "*Substitution*" function and we denote each of our selections and use the seperator that we want
`$3-$1-$2`

(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)

The REGEX' are saved and can be accesed through the links in the task.

From Voyant to R (HTTPS://REGEX101.COM/R/0ZOIKA/1)

To get from the stopwordlist that is meant for Voyant to R, we first write the following:
`(\S+)\n?`

Then we write the following to substitute (in regex101.com):
`"$1",`

From R to VOYANT (https://regex101.com/r/nG1nJ6/1)

To get from the stopwordlist from R to Voyant, we first write the following to match all words. Here we have decided to select all words between the quotation marks. /S matches anything BUT whitespaces (i.e., words). We catch the last word with the "?"
`"(\S+)",?`

Then we write the following (in regex101.com) to substitute (using the Substution function).
`$1\n`

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

Firstly, variables should be in columns and only one value in each cell is a great principle. Having multiple values in each cell will only create confusion and mess up the data analysis. Each observation should be one row (e.g., one row for each participant in an experiment so that all of the data to that participant is connected).

Being consistent is highly important for good data organisation in spreadsheets. For instance, if you want to record gender identity, recording it as "M", "MALE", "MAN" will give you extra work when doing preprocesing. Rather, one should settle on one way of denoting the value (e.g., "M", "F" or "X"). Consistency is also important when considering missing data. NA can be used to denote a missing value (abbreivation for "not available"). This is more transparant than having an empty cell which may signify other errors in data recording.

If the plan is to save to a CSV file, it is important to only make one sheet as CSV only saves the active sheet.

Finally, one should aim to keep track of the meta data of the data that is collected. This is especially the case for more abstract variables that may be very specific to the project that you are working on. Naturally, a variable such as *nationality* may be self-explanatory but a more abstract variable needs explanation if it is to be used in the future where the data collectors may not remember the details of the project.

Reference: https://datacarpentry.org/spreadsheets-socialsci/01-format-data/index.html