Regular expressions and spreadsheets homework 1

Joshua Hansen

2022-08-30

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 ?

Answer to 1): https://regex101.com/r/um8qMj/1

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)

From Voyant to List: https://regex101.com/r/tUR6bA/1

From List to Voyant: https://regex101.com/r/tUR6bA/2

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

Answer 3): A good practice would first of all to be consistent with the data. You should always try to follow through with any conventions you have so you can manipulate all data entries at once. That means that for each column there should only be one data type and one way of spelling. You should also try to make the data entries computer-readable for further analysis. Colors are e.g., not readable in R and should be conveyed differently.

You should always have a documentary on each column or entry on what the data type is, what it means and describe the context of it. You could resolve some issues by implementing a data validation, especially when collaborating with others. One Column should also only have one value or information and each entry should have an individual keyID. It is also good to not have multiple pages in one sheet as the computer will only export a single page.

In general you should try to make every data entry understandable for a human and provide intuitive description and documentations of your data. Spreadsheets are good for having an overview of your data or making simple statistics, but they lack versioning and are prone to invalid data entries. It is always recommended to have a changelog that will show any changes made in the data, as raw, captured data will be very valuable and necessary for a complete research.

4. 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)!

4) next time 😊