# Homework w36

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 ?***

Solution: <https://regex101.com/r/qaYZqy/1>

\s? is used to make sure that it also matches with the dates containing a space

Parentheses divides the different parts of the date

Using $ to refer to the different parts of the date, it’s possible to rearrange in the preferred order

1. ***Write a regular expression to convert the stopwordlist (list of most frequent Danish words) from Voyant in***[***http://bit.ly/reg toexexercise3***](http://bit.ly/reg%20toexexercise3)***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)***

**Voyant 🡪 R: Solution:**

<https://regex101.com/r/KwmqC1/1>

The stopwordlist in Voyant-format

The expression *\n* matches with the newline between the rows and gets replaced with *“, “* to fit the R-format

**R🡪Voyant: solution**

<https://regex101.com/r/cR3THJ/1>

*“, “* gets replaced by the expression, *\n,* which makes a new line and transforms it into Voyant-format

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

Some principles for using spreadsheets for good data organization, as addressed in "Data Organization in Spreadsheets" by Karl W. Broman and Kara H. Woo, are crucial for maintaining data integrity, accuracy, and usability, and can be summarized into the following points

* + **Use consistent formatting:** Data has to be consistently formatted within columns. Use a single format for dates, numbers, and text.
  + **Organize data hierarchically:** Use headlines and sub-headers to arrange data hierarchically to create a logical structure. This ensures easy navigation and understanding of the data
  + **Avoid merged cells:** Merged cells can lead to confusion and hinder sorting and filtering. Keep cells unmerged to maintain data integrity.
  + **No blank rows or columns:** Get rid of unnecessary blank rows and columns. They can cause problems in data analysis and lead to misinterpretations.
  + **Use descriptive column headers:** Choose precise column headers that describes the content accurately. This helps users understand the data and perform meaningful analysis.
  + **Ensure data consistency:** Enforce data consistency through drop-down menus, data validation, or controlled vocabularies. This minimizes errors and enhances data quality.
  + **Separation of data entry and data storage:** Keep raw data separate from summary or calculated data. Avoid to put calculated values in the same cells as raw data.
  + **Document the spreadsheet:** Provide detailed documentation with explanations of data sources, definitions of column headers, and any data changes or calculations.
  + **Version control:** Use version control to track changes made to the spreadsheet over time. This helps identifying errors.
  + **Data validation:** Validate data for errors at regular intervals, inconsistencies, and outliers. Use built-in Excel functions or external tools to identify and address issues.
  + **Backup and share responsibly:** Regularly backup your spreadsheet to prevent data loss, and share it securely with only authorized individuals to maintain data confidentiality.

By adhering to these basic principles, users can create well-organized spreadsheets that are more reliable, accessible, and conducive to effective data analysis and decision-making.

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!.*

/\b[Dd]i+s.[Mm]anibus.[Ss]acrum\b|\b[Dd]i+s.[Mm]anibus\b|\bD\s?M\s?S\b|\bD\s?M\b/gm

Et billede, der indeholder tekst, elektronik, skærmbillede, software

Automatisk genereret beskrivelse

Solution, but only with the first 20 rows (it was unable to save with all the text)

<https://regex101.com/r/owmcHp/1>