TU Dortmund University
Department of Statistics
Dr. J. Rieger
M. Sc. K.-R. Lange

Winter Semester 2023/24
10/30/2023

Exercise sheet 2

Text as Data

Hand-in (voluntarily): 11/03/2023 until 10:00 a.m. via Moodle
Please submit a `.py`, `.ipynb`, `.R` or `.rmd` file!

Task 1

In Moodle you will find the file `110723EUParl.docx`. Read the text inside into your console.

It contains a plenary protocol of the European Parliament from the 11th of July 2023. In this exercise, you will use regular expressions to extract meta information from the file and separate the text into smaller parts. Write functions that can be generalized to other protocols of a similar structure. That is: when you are for instance trying to remove the table of contents, do not just remove the first 7 pages, but find a way to automatically detect when the main part starts (for instance using Regex).
When you want to find a certain set of tokens using Regex you might want to proceed as follows:

1. Look at the original document and identify structural features that could be used to identify the tokens you are looking for.
2. Translate these loose strucutral features into Regex.
3. Use Regex to see whether you detect your desired tokens **and only your desired tokens**.
4. Modify your Regex to solve the exercise.

Task 2

Identify the date and the weekday of the plenary discussion. Transform the date into a date-object that can be used to create timelines in your programming language (e.g. package `datetime` in Python and `as.Date()` in R).

Task 3

From your large text, filter out the Attendance Register (here: page 35) and the cover sheet as well as the table of contents (here: everything until page 8) and remove them so that only the main part of the document is left.

Task 4

Split the text into the individual chapters provided in the original document.

Recommended packages & functions

R: `gsub(), stringi::stri_extract_first, stringi::stri_detect_regex, officer::docx_summary(), officer::read_doccx()`
Python: `re.search(), datetime.datetime.strptime(), docx.Document()`