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

Winter Semester 2023/24
10/23/2023

Exercise sheet 1

Text as Data

Hand-in (voluntarily): 10/27/2023 until 10:00 a.m. via Moodle
Please submit a `.py`, `.ipynb`, `.R` or `.rmd` file!

Task 1

In Moodle you will find three files, each containing 2 movie reviews: `reviews1.txt`, `reviews2.txt` and `reviews3.txt`. One of the files has a UTF-16 encoding, while the other two are UTF-8 encoded. Check them for their encoding and load the texts within them into your console.

Task 2

Split the lines in all texts (seperator "\n") to separate the two reviews in each file and then combine the reviews from all three files into one list. The result should thus be a list of six strings.

Task 3

Apply elementary text handling ("preprocessing") steps. That is, within each review

- Remove punctuation, numbers and special characters
- Turn all letters into lower case
- Split the text into individual words

The result should be a list of lists of Strings (Python) or a list of character vectors (R). Each inner list/character vector represents a review as separated words.

Task 4

Count how often each word occurs in this text corpus and display the 5 most common words. Can we use these top words to infer what the texts are about?

Recommended packages & functions

R: `gsub()`, `strsplit()`, `tolower()`, `table()`, `tm::removePunctuation()`, `tm::removeNumbers()`, `readLines(file(...))`
Python: `str.isalpha()`, `str.isspace()`, `str.split()`, `str.lower()`, `collections.Counter()`, `open(...).readlines()`