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

Winter Semester 2024/25
10/15/2023

Exercise sheet 1

Natural Language Processing

Hand-in (voluntarily): 10/21/2024 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 `magic.txt`. Load the data inside into your console.

It contains the description of 1419 *Magic: The Gathering* cards. This description contains crucial information about each card, such as its name, its mana cost, its type and its effect. Your task will be to extract all crucial information from this data set of unstructured text and turn it into a well-structured data format.

Task 2

Each line in the document represents the information about one card. Split the lines (separator "\n") to be able to look at each card individually. The result should be a list of strings/a character vector.

Task 3

The information about each card is given in the following format:
`CardName: [...] CardCost: [...] CardType: [...] CardEffect: [...]`
Exploit this format to extract and save each bit of information separately. Turn the information you collected into a coherent data frame with the columns "Name", "Cost", "Type" and "Effect".

Task 4

Which of the words "Creature", "Sorcery", "Instant", "Enchantment" and "Artifact" appears most often in all the texts within the "Type" column?

Recommended packages & functions

R: `strsplit(), table(), readLines(file(...)), data.frame(), grepl()`
Python: `str.split(), collections.Counter(), open(...).readlines(), pandas.DataFrame()`