Technische Universität Dortmund
Department of Statistics
Dr. J. Rieger
M. Sc. K.-R. Lange

Winter Semester 2022/23
11/02/2022

Exercise sheet 2

Text as Data

Hand-in (voluntarily): 11/06/2022 until 11:59 p.m. via Moodle

---

Task 1

In Moodle you will find the file `Potter.zip`. Unpack it. It contains 7 txt-files, each containing the text one of the "Harry Potter"-books. Load those txt-files into your programming language.

Task 2

To compare the books, we must know, which book it is we are looking at. Each file contains one particular line for every page in the book:

Page | `page_number book_name` - J.K. Rowling

Use regular expressions to automatically detect the name of the book from the texts.

Task 3

The texts in the the txt-files are not "clean" yet. To analyze them properly, we need to do additional preprocessing steps.

- Remove the page indicator from the texts. That is, remove all lines that have the form mentioned in task 2
- Trim the start of the document until the first chapter starts.
- Remove the headers of all chapters. These are written in CAPS (all letters are capitalized). Detect this using regular expressions.
- Replace all line breaks ("\n") with a whitespace (" ").

The result should be a list of 7 large stings, one for each book.

Task 4

Just like in the first exercise sheet, apply elementary preprocessing and tokenization steps. Afterwards apply lemmatization and stop word removal. The result should be a list of lists. Each inner list represents a book as a list of words.

Task 5

Calculate the tfidf for your corpus. Return the words with the highest tfidf for each of the 7 books. Does the result give you an idea of what the books are about? If not, why?

Recommended packages & functions

R: `stringi::stri_match()`, `tidytext::bind_tf_idf()`
Python: `re`, `sklearn.feature_extraction.text.TfidfVectorizer`