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

Winter Semester 2022/23
10/26/2022

Exercise sheet 1

Text as Data

Hand-in (voluntarily): 10/30/2022 until 11:59 p.m. via Moodle

---

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.
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 2

Apply elementary tokenization steps. That is, within each review

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

The result should be a list of lists (list of vectors for R). Each inner list represents a review as a list of words.
Count how often each word occurs in this text corpus and display the 5 most common words.

Task 3

Use each one automated word stemming- and lemmatization method for your programming language. Apply them to the corpus resulting from task 2 and compare the resulting texts when applying each. Which of the two approaches would you prefer?

Task 4

Use your "best" corpus from task 3 and apply stop word removal. That is, remove every word from a stop word list from your text. Beware that you have to apply the same pre-processing of your text to your stop words, such as removing the apostrophe from "don't".
Compare the most common words with the results from task 2. What do you notice?

Recommended packages & functions

R: `gsub()`, `strsplit()`, `tolower()`, `table()`, `tm::removePunctuation()`, `tm::removeNumbers()`, `tm::stemDocument()`, `tm::stopwords()`, `textstem:lemmatize_words()`
Python: `str.isalpha()`, `str.isspace()`, `str.split()`, `str.lower()` `collections.Counter`, `nltk.stemmer`, `nltk.corpus.stopwords`