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

Winter Semester 2024/25
10/22/2024

Exercise sheet 2

Natural Language Processing

Hand-in (voluntarily): 10/28/2024 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. Load the texts within them into your console. If you have used the correct encoding, the texts in your console should be readable for a human. Each review should be one element in a list of six total elements.

Task 2

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.
Count how often each word occurs in this text corpus and display the 10 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 preferred 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()`, `stringi::stri_replace_all()`, `tm::removePunctuation()`, `tm::removeNumbers()`, `tolower()`, `tm::stemDocument()`, `tm::stopwords()`, `textstem:lemmatize_words()`
Python: `str.isalpha()`, `str.isspace()`, `re.sub`, `str.lower()`, `nltk.stem.PorterStemmer`, `nltk.stem.WordNetLemmatizer`, `nltk.corpus.stopwords`