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

Winter Semester 2023/24
06/11/2023

Exercise sheet 3

Text as Data

Hand-in (voluntarily): 11/10/2023 until 10:00 a.m. via Moodle

---

Task 1

In Moodle you will find the file `trump.xml`, containing Speeches of Donald Tump's speeches during the campaign rallies of his 2016 presidential election.
The file also contains meta information about the place and date of the speech. We are however only interested in the speeches themselves. Read the xml file into your console as if it were a simple text-file and then use Regex to filter out the speeches.

Task 2

Apply elementary tokenization steps. That is, within each speech

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

The result should be a list of lists (list of vectors for R). Each inner list represents a speech as a list of 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 "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: `tm::stemDocument()`, `tm::stopwords()`, `textstem:lemmatize_words()`
Python: `nltk.stem.PorterStemmer`, `nltk.stem.WordNetLemmatizer`, `nltk.corpus.stopwords`