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

Winter Semester 2022/23
11/09/2022

Exercise sheet 3

Text as Data

Hand-in: 11/13/2022 until 11:59 p.m. via Moodle

Task 1

In moodle you will find the file `hashtag_donaldtrump.csv` and `hashtag_joebiden.csv`. Load the texts within them into your console. They contain tweets with the hashtags `#DonaldTrump` and `#JoeBiden` bevore the American presidential campaign in 2020. Filter the data for tweets within the Country "United States of America".
We are interested in the columns "tweet", "created at" and "user description".

Task 2

Take at the look at the texts and use your knowledge about Twitter-specific-vocabulary and common components of Tweets to preprocess the texts so that they are fit for an anaysis.

Task 3

In Moodle you will also find the file "Afinn.txt". It contains a sentiment lexicon which yields a sentiment value for a large amount of words. Use this lexicon to calculate the sentiment score of each tweet.
Note: use an efficient data structure (e.g. hash in R, dict in Python) to store the lexicon. The sentiment analysis will take much longer otherwise.

Task 4

Filter the tweets for users whose description contains either the words "democrat" or "republican". Does the average sentiment between both user groups differ for one of the hashtags?

Task 5

Use your result from Task 3 and the column "Created at" from the original files to create a timeline of sentiment. How does the sentiment of tweets containing either hashtag over time? Display the average sentiment over time visually.

Recommended packages & functions

R: hash, read.delim
Python: pandas, matplotlib, datetime