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

Winter Semester 2023/24
11/20/2023

Exercise sheet 5

Text as Data

Hand-in (voluntarily): 11/24/2023 until 10:00 a.m. via Moodle

---

Task 1

In Moodle you will find the file `bitcoin.csv`, containing Reddit comments of the bitcoin-Subreddit from 2022. Read the file into your console.
From this data set, we only need the columns "created", determining the date at which the post was created, "title" containing the title of the post and "selftext" containing additional text from the post, if any. For our analysis, we want to analyze "title" and "selftext" as one combined entitiy for each text. So for each post, join the two respective strings if there is a selftext.

We will now perform a simple sentiment analysis and compare the resulting time series with the actual Bitcoin price, which you can find in `bitcoin_prices.csv`.

Task 2

Apply preprocessing to the given texts. Keep in mind, that we intend to use sentiment dictionaries to analyze the text later. How does this knowledge change your approach to preprocessing?

Task 3

In `dictionary.csv` you will find a sentiment dictionary. "Positive words" will have positive values while "Negative words" will have negative values.
Use this dictionary to calculate the sentiment score of each text, that is sum up all sentiment values to the corresponding words in said text. A negative score will thus indicate a negative text, while a positive value will indicate a positive text.

Task 4

Compare the daily difference in market values in the file `bitcoin_price.csv` and your sentiment scores with a correlation coefficient of your choice. Do the comments explain the behaviour of the bitcoin price evolution well?

Recommended packages & functions

R: `corr()`, `as.POSIXct()`, `tidyverse::group_by()`, `tidyverse::summarise()`
Python: `pandas.to_datetime()`, `pandas.groupby()`, `pandas.corr()`