Data Transformation with Tweets

Practice Exercise

Prof. Dr. Nicolas Meseth

1. Simple Transformations

Try to solve the following problems using R and the Tidyverse to practice your data transformation skills.

1. Filter the data set for tweets that are not retweets (i.e., `is_retweet` is FALSE).
2. Extract the top 10 tweets with the highest `retweet_count`.
3. Create a new column called `is_reply`, where the value is `TRUE` if `in_reply_to_screen_name` is not `NA` and `FALSE` otherwise
4. Filter the data set for tweets with at least one hash tag. Sort the tweets by number of hash tags in them.
5. Filter the data set for tweets in the English language (`lang == "en"`) and containing at least one URL.
6. *Unnest* the `urls` column. Then, count the frequency of each unique domain (e.g., "twitter.com").
7. Create a new column called `created_day` which extracts the date (without time) from the `created_at column`. Then, count the number of tweets per day and visualize the result as a time series.
8. Identify tweets that mention other users. *Unnest* the `user_mentions` column and calculate the number of times each user is mentioned. Then, find the top 10 most mentioned users.
9. For each user, calculate the proportion of their tweets that are retweets, replies, and quotes. Then, visualize the results as a stacked bar chart, where the x-axis represents users and the y-axis represents the proportion of each type of tweet.

2. Food for Thought

Try to answer the following questions to deepen your understanding of the topics around data transformation with R and the Tidyverse.

- What are advantages of a Tibble over the classic R data frame?
- What are the benefits of using the pipe operator `|>` when working with data in R? How does it improve code readability and organization?
- Research and explain the concept of "laziness" in data manipulation with dplyr. How do "lazy" operations improve efficiency when working with large datasets?