Data Transformation and Visualization

Mandatory Exercise

Prof. Dr. Nicolas Meseth

OVERVIEW

To develop proficiency in natural language processing (NLP), it is essential to first update your general data exploration skills. For this reason, your initial exercise will not involve working with unstructured data. Instead, you must demonstrate your ability to use **the five fundamental data transformation techniques** and employ **the basics of data visualization** with structured data first.

Here are the relevant lessons from the online script. Make sure you study them carefully, including the provided literature, in order to solve this exercise:

- 1 - Working Environment
- 2 - R and the Tidyverse
- 3 - Data Loading
- 4 - Data Transformation
- 5 - Data Visualization

Good Luck!

YOUR TASKS

To pass this exercise, you must complete the following tasks and submit your results via ILIAS. You find the submission details below.

Task 1: Install the working environment

In this task, install the necessary software for the development environment on your computer and make sure everything works properly. You find all instructions and download links in the online script. As a reminder, here is a list of the software you need to install:

- A current version of the R programming language (>= 4.2.3)

- The latest version of RStudio Desktop
- A current version of the Python programming language
- Git

To test your working environment, open a new terminal in RStudio and checkout the GitHub code repository for the WordLens project. Install all required R-libraries and run the first script `1_why_data_transformation.R` from top to bottom. If everything works properly, the script should produce two different visualizations.

Task 2: Transforming tweets

In this task, you must apply your data transformation skills to modify the `tweets` data set in various ways. The goal is always to create a view on the data that gives answers to questions or that allows visualizing the data afterward.

1. Create a transformation to get only the original tweets from Anna-Lena Baerbock!
2. On average, which user gets the most likes (favorites) for their original tweets?
3. Calculate the percentage of retweets vs. original tweets for each user in the data set!
4. Create a new Tibble with the months of 2022 in one and the number of tweets in another column! Group the data by user!
5. On average, which users in the data set add the most hash tags to their tweets?
6. What is the average tweet length for each user in the data set?
7. Create a top 10 list showing who writes the longest tweets on average! Consider only original tweets!
8. What hour of the day is the most active regarding the number of tweets? Create a tibble sorted by the hour of day and a second column with the average number of tweets!

Task 3: Visualizing tweets

In this task, you must combine your data transformation and data visualization skills to create a plot for each of the following questions:

1. Is there an association between the number of retweets and the number of likes (favorites) a tweet gets?
2. Visualize the development of the total number of tweets per month by political party (SPD, Grüne, FDP)! What do you observe, and how could you explain this phenomenon?
3. Create a suitable visualization for the last question from the previous exercise to uncover patterns in the use of Twitter over the hours of the day!

4. Extend on the previous visualization and create a separate plot for each political party (SPD, Grüne, FDP) in one visualization (hint: `facet_wrap` might help) to highlight potential differences in usage patterns across the political parties.

SUBMISSION

For your first exercise, please submit:

- Two R-scripts with your solution to the tasks 2 & 3:
    1. `task_1_transforming_tweets.R`
    2. `task_2_visualizing_tweets.R`

Submit all files via the corresponding exercise in ILIAS.