

## Analyzing the Rotten Tomatoes Movie Review Dataset

This tutorial concerns with handling a tabular string dataset. We will showcase how to load & explore the data, conduct a linear regression and perform a textual analysis.

### Exercise 1: Data Loading and Exploration

- Download the Rotten Tomatoes Movie Dataset from <https://openBIGdata.org> load the data to memory (e.g., using pandas in Python).
- Identify the number of variables and observations in each of the two sub-datasets.
- Explore the distribution of movie genres in the dataset. To do so, identify the most common genres and visualize their frequency in a bar plot.
- Identify any missing values in the two subdatasets and justify how you would deal with them in your subsequent empirical analyses.
- Compute summary statistics for the numerical variables in the dataset.
- Visualize the distribution of movie release dates over time in an appropriate plot. At which point have been the most movies released?

### Exercise 2: Regression Analysis

This exercise is to understand how variables in our dataset are related to each other.

- Perform a linear regression to predict audience ratings based on tomatometer ratings. Interpret the results.
- How are movie runtime and tomatometer ratings related? Create a scatter plot with a trend line to visualize the relationship. Compute the correlation coefficient between the two variables.
- Predict the likelihood of a movie being rated as “Fresh” using a logistic regression. Which predictors would you choose to do so? Why? Interpret the outcomes of your model.

### Exercise 3: Textual Analysis

Besides numerical ratings, the Rotten Tomatoes Movie Review Dataset also contains review texts about movies. In this exercise, we will learn how to analyze these texts using various methods.

- Load the reviews file.
- Pre-process the review texts by removing punctuation, converting the text to lowercase, and tokenizing the text.
- Conduct a sentiment analysis on the pre-processed textual reviews, using, e.g., the Afinc dictionary as to obtain an average sentiment of each movie review in the dataset.
- Average the sentiment scores across movies. Which movie is rated best according to its average review sentiment score?
- (Advanced) Investigate the relation of movie genres. One movie can belong to multiple genres. Create a network graph that shows which genres are mixed.
- Use ChatGPT or an alternative AI tool (see e.g., <https://www.gettingstufdone.ai/>) to replicate the code for the sub questions in this task. Compare your code with the computer-generated code. How would you assess its quality relative to yours?