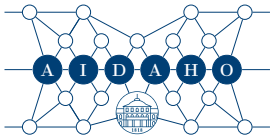
**Introduction to Applied Data Science**
Prof. Dr. Thomas Dimpfl, Dr. Johannes Bleher, & Sophia Koch
Department of Business Mathematics
and Data Science




# Mandatory Assignment 2

—

# Sentiment Analysis of User Reviews

Before starting the asssignment, thoroughly review the document **"Assignment_Guidelines"** that you can find on ILIAS. To submit your assignment, store your R script(s), pdf-file, and your individual dataset(s) in a Git repository. Add Prof. Dr. Thomas Dimpfl (dimpflth), Dr. Johannes Bleher (jbleher), and Sophia Koch (kochsophia) as members (role: maintainer) to your Git repository before the assignment deadline expires. Additionally, submit **only** your written report on ILIAS (See Section 4. Upload in "Assignment_Guidelines"). The Assignment deadline is **January 7$^{th}$, 2024, 11.55 p.m.**.

## Task 1: Initiate Your Git Repository

1. Initiate a git repository locally and on the AIDAHO-Gitlab. Call the repository `IntroADS_Ass2_TeamYY` *(YY should equal your team number from ILIAS).*

2. Generate the following folder structure (if you like you can adjust the shell script from the lecture video accordingly, available on ILIAS):

```
IntroADS_Ass2_TeamYY
├── 00_docs
│   └── literature
├── 01_data
│   └── raw
├── 02_code
│   └── R
└── 03_report
    └── graphs
```

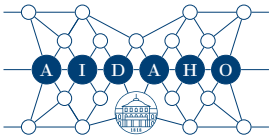
3. Save the assignment in `00_docs`.

4. Copy the data that you have collected in Assignment 1 into `01_data/raw`.
   ***IMPORTANT:*** *The dataset `gamereviews.RData` should contain the reviews from requesting 100 reviews 100 times. Thus, the dataset has either 10.000 observations or the maximum number of reviews available for the game.*

5. All your R scripts should be saved in `02_code/R`.

6. Your written report should be saved in `03_report`.

7. All graphs, that you display in your report, should be additionally saved in `03_report/graphs`.




## Task 2: Clean your Game Reviews

Clean your data set `gamereviews.RData`. You have learned about several cleaning steps in Exercise Sheet 6. Cleaning procedures that you must perform:

- Drop all non-ASCII characters.
- Transform all abbreviations "n't" into the actual word "not".
- Delete all stopwords except for the word "not".
- Exclude all observations with empty reviews from your dataset.

Save the cleaned dataset in `01_data`.

***Hint:*** *Split your code into two R scripts. One should perform the cleaning tasks and the other one should contain the code for the following analysis.*

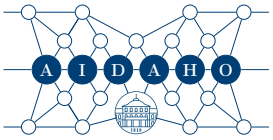
## Task 3: Sentiment Analysis of your Game Reviews

The main part of this mandatory assignment is a sentiment analysis of the game reviews.

We expect two analyses: First, an analysis implemented manually using the Bing sentiment word list presented in the lecture. Follow the steps on slide 24 of lecture 8 ("VL_08_TextAsData.pdf"). The necessary word lists of positive and negative words can be downloaded from Kaggel. Second, we expect a deeper analysis using the `syuzhet` package. The package also contains the Bing dictionary. Compare the sentiment score you get from your basic analysis with the result using the `syuzhet` package. If you don't get identical results, open the black box and check what the package function does. Use two more dictionaries and the possibility to also analyze emotions. Wordclouds are nice representations, but appropriate measures might be more suitable for a formal analysis.

Thereby we expect you to answer the following questions:

- Is the sentiment score and the allocation of being a positive or negative review consistent?
- Are reviews with extreme sentiment score (high or low) seen as more helpful compared to reviews with a moderate sentiment score?
- Does the average sentiment score resemble the over all review score of your game?
- Does the sentiment vary over time?




If on the way you find that further questions arise, present them in your report and answer them. This is data science, questions come up on the fly while working with data. So the above presented questions are really a rough guideline, it is up to you to make it nice and worthwhile.

As there is probably more to say than in the first assignment, we will raise the page limit to **6 pages**.