Introduction to Data Science, WMCS16002, semester 1b 2016

1 Homework

Due November 22, 2016 12:00:00 (noon) CET
Submit by creating a pull request on GitHub

Your submission should consist of

- A brief written report (preferably PDF or dynamic report formats (RMarkdown, etc.) but OpenOffice is fine as well
- Source code files which generate the solution, tables, visualizations used in the report. The source code can contain much more than what is finally used in the report—please use comments to structure the source code.

Note that you are free to use whatever programming language you want if not stated otherwise in the assignment.

You have been allocated to work in groups of three to four people to mix different backgrounds. This is an interdisciplinary course therefore we suggest you to take advantage of your complementary backgorund (including Mathematics, Astronomie, Engeneering and Computing Science). Please indicate in the written report who your group members are and also who did contribute to what.

Remember that you have to pass every homework assignments (but one) to pass the course. Furthermore, note that **plagiarism is fraud** and we take it serious. If we find it in your submissions you risk being expelled.

Submission

- The bonus parts are still sent with the pull request, but remember to report on who contributed to what.
- Make a new folder for every assignment (this is the first one). Every assignment's folder should have clear instructions on how to run your code.
- Do not commit massive data files to the repository, but leave clear instructions on what goes where.

Good luck and have fun working on the exercises!

Before Anything Else

1. Read up on git and GitHub[1] if you are not already familiar with distributed version control.
2. Create a GitHub account.
3. Join the repository of your team (`https://github.com/RUG-IDS/team-xx`). We can do this for you when give us your GitHub username.
4. Clone your repository's `dev` branch and prepare your assignments here.
   `git clone -b dev git@github.com:RUG-IDS/team-xx.git`[2]

Note: Only **we** have push access to the `master` branch. You work on the `dev` branch. Each of the assignments should be in a separate folder. When ready, you submit your work by sending a pull request to the `master` branch of your team's repository. We assess the work that is submitted through the pull request.

Hollywood Data Science

You might have encountered the Internet Movie Database IMDb `http://www.imdb.com`, which is a collection of film related information. Most of the data can be downloaded as plain text files and some tools are available allowing to search and display information.

Download the file `movievalue.csv` we provided in Nestor. This file contains limited information about the movies. To obtain more information there are two ways:

FTP The FTP server of Freie Universität Berlin (Germany) contains a subset of IMDb data of manageable size. From `ftp://ftp.fu-berlin.de/pub/misc/movies/database/` We also provide an `imdb-data-parser-master.tar.gz` which you might use for the data import if you like.

API The data can be obtained also from the OMDB API (`http://www.omdbapi.com/`). However, there you have to request information for every movie individually.

1.1 Collect it . . . link it! (60P)

Collect and clean the data about the movies in the `movievalue.csv`. Enrich that data collection by acquiring also the `Genre, imdbRating, imdbVotes` (and optional also Director, Country, PG rating, etc.) for at least 1000 of the given movies for further analysis.

TIP: One of the two ways of getting the data (FTP vs. API) is easier than the other. You have to figure out which.

If you are using R, Matlab or Python, the following might be useful:

1. `data.frame` (R)
2. `table` (Matlab from R2013b)

[1] A good source is `https://help.github.com/articles/about-pull-requests/`

[2] Or if you want to use HTTPS: `git clone -b dev https://github.com/RUG-IDS/team-xx.git`.

3. DataFrame (Python).

Note, linking the data might be tricky since names might not be unique (episodes of series often reuse names of movies and pre/sequels may be written differntly)!

1.2 Types (20P):

The data set(s) you constructed in 1.1 contains features like “ReleaseDate”, “Movie”, “Production Budget”, “Domestic Gross”, “Worldwide Gross”, “genre”, “imdb rating”, “number of imdb votes”, “director”, etc.
Determine the data type of each of them and explain your decision shortly.

1.3 Data Science Projects (20P):

a) Think about at least one descriptive question on the data set and present it by using a table, scatter plot, box plot and/or histogram.

b) Think about at least one exploratory analysis on the data set and present it by using a scatter plot, box plot or histogram.

1.4 Bonus (+20P):

Acquire ratings from `www.rottentomatoes.com` and compare them with the IMDB ratings. What do you observe? How and why are they different?

Do you find evidence that the appearance of a certain actor in a film leads to better ratings or box office success than others? Or does the country of production have a bigger impact on the financial success? Does a bigger production budget generally lead to more success in the box office?