Data Cleaning for Movies

It might sound a bit abrupt, but clean data is a myth. If your data is dirty, so is everyone else's. Enterprises are more than dependent on data these days, and it is going to stay the same in coming years. They need to collect data in order to analyze it, which necessarily will not be 100% clean, pristine, or perfect in nature.

Nearly all companies face the challenge of dirty data in the form of a lot of duplicates, incorrect fields, and missing values. This happens due to omnichannel data influx, followed by hundreds, if not thousands, of employees wrestling and torturing that data to derive professional outcomes and insights.

The goal of this project is to clean and make the data ready for downstream analysis. But the exact recipe making dirty data to the process one also matters, which will be shown in the project.

The steps of the project

- Appropriately labels the data set with descriptive variable names.
- Extracts only the measurements with impact on analysis.

- Merges two test sets to create the one using ID.

Dependencies

- R version 3.5

- install `install.packages('data.table')`

- install `install.packages('dplyr')` which is necessary to cleaning data

how to use

- Oroginal data sets are `movie_info.tsv` and `reviews.tsv`.

- The codebook is in `codebook.md`. It gives the descriptions of the variables in the data frame prouduced by this project.

- Please Run `Run.R` to see R codes producing tidy data.

- The final tidy data is in `tidydata.txt`. It can be loaded by `read.table("tidydata.txt", sep= " ", header= TRUE)`.