Project 1 part 2 report

Yapi Donatien Achou

October 11, 2020

1 Introduction

The objective of this part of the project, was to clean the data. The original data came as three separate files. A json file containing users data, an excel file containing the films information such as title and genre, and a dat file containing information about the users ranking.

2 Data cleaning process

The users data contained in the json file was uploaded into a pandas data frame by using the read json pandas function. After inspecting the types of each columns, the entries containing None and NaN were removed. The resulting data frame was saved as a csv file.

The dat file containing the ranking information was loaded into a pandas data frame, inspected and missing values represented by None or NaN were removed. The data frame was subsequently saved as a csv file.

The Excel file was loaded into a pandas data frame. The genre were added as columns, to form a new data frame, which was saved as a csv file.

The data cleaning process was implemented in three separate functions, each dealing with each file (json, dat and excel). A wrapper function was implemented to save the csv files into an output folder, which is one of the arguments of the wrapper function, the second being the name of the folder containing the raw data.

3 Conclusion

In the final stage of the project, the three csv files (users, rankings and films), will be merged into a single file, which will represent the final data for constructing the final model for the movie recommendation system.