Matt Hill

LIS 545

January 27, 2023

Data and Metadata Profile - [The Movies Dataset](https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/versions/5?resource=download)

For this assignment I have chosen to review “The Movies Dataset'' from kaggle.com, which is a collection of files related to movies released prior to August 2017. There are six csv files in total that include a keywords file for plot keywords, a credits file for cast and crew information, a ratings files for movies ratings, a couple files that contain movie ID information for TMBD and IMDB, and a movie metadata file for all extraneous information related to the movies like posters, budget, release dates, and other things. As long as you have a program like Excel to open the csv files you will be able to look through them. There are no additional holds from accessing this data. This data was collected from the Full MovieLens Dataset and includes ratings data from the official GroupLens website.

This data set was created by Rounak Banik for his Capstone Project. There are a lot of interesting things people could do with this data like review box office information, get a general understanding for how well any given movie was received by audiences, search for trends around movie performance and release dates, or other statistical information someone might be curious about. I think the people who would most likely use this information are general movie fans, movie production companies, film crews, or researchers who are interested in trends around film.

The metadata involved with this dataset is found within the files themselves and are mainly categories to organize the existing data. While there is a file in the dataset labeled “metadata”, this file does not actually store metadata about the dataset in general, but includes information about the movies included in the dataset that is not captured in other files. There is some additional information about the dataset on kaggle.com to help add context and descriptions to what can be found within these files, but I wish this information went into a bit more detail as it doesn’t really explain how to read the dataset in any way.

I think the most obvious way this dataset can be improved is by updating it with movies created during or after August 2017. Aside from this, I think it would be very helpful for the creator of the dataset to consolidate their work into one place on kaggle. There are three separate notebooks this author created that were all related to this same dataset. It seems he made one notebook to review the dataset, another specifically for a movie recommendation system, and then another that was a summary of all of this work. This was really confusing to me because I initially thought there were different datasets for each notebook and couldn’t initially see differences between the files that were on each page. I also think that, while focusing on the budgets and financial success of these movies is interesting, there are many other ways that a film can be considered successful aside from box office performance. It would be really neat to see information related to awards given to each movie and compare that to their box office performance. Timelines for how long a movie was playing in theaters could also be an interesting metric to look at when considering a film’s financial success. For example, The Rocky Horror Picture show was a financial failure when it was initially released, but it has also become one of the longest theatrical runs of a movie in theaters as it has been played consistently in theaters since 1975. I think the analysis that has been run on this dataset was a great first pass, but there is so much more that can be done.

I could not find any publications that referenced this dataset in their work. There were no listings of publications that could be found on the dataset’s webpage on kaggle. I also tried googling the author’s name, the dataset name, and a combination of these but could only find links back to the same page on kaggle.

Overall The Movies Dataset provided an interesting collection of information for a wide pool of films throughout history. While the author mainly talked about how this data could be used to help predict financial success for movies, there are many more things to be done with this data. The metadata related to this information is pretty straight forward in that it mainly describes the type of information related to a given film like title, language, genre, and runtime but there could have been a lot more helpful metadata involved. Especially for folks who are simply coming across these files and may not know exactly what’s in them. It would be nice to have a true metadata file that helps researchers understand where they can find specific types of data without having to open each one.

References

Banik, R. (2017, November 6). *Movie Recommender Systems*. Kaggle. Retrieved January 29, 2023, from https://www.kaggle.com/code/rounakbanik/movie-recommender-systems/notebook

Banik, R. (2017, October 24). *The movies dataset*. Kaggle. Retrieved January 29, 2023, from https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/versions/5?resource=download

Banik, R. (2017, October 25). *The story of film*. Kaggle. Retrieved January 29, 2023, from https://www.kaggle.com/code/rounakbanik/the-story-of-film/notebook