Kayla Crosbie

Final Project

University of Denver

February 25th, 2021

**GIThub link:** https://github.com/kcrosbie93/FinalProject

**Dataset**

The selected dataset contains data of all the tv shows and movies on Netflix, as of 2019, and remains updated monthly since initial compilation. The dataset was created and managed by Shivam Bansal. The data is gathered from Fixable.com, a third-party search engine for Netflix through regular API calls. Stakeholders of this data would include Netflix creators, Netflix users, international users, other video hosting sites, movie and tv enthusiasts, researchers, and the community of data scientists utilizing Kaggle.

Kaggle allows users to upload their data, and others are able to interact with the data, pose challenges to use the data to solve problems, and share python codes on a cloud-based workbench. Often, Kaggle users will use data they have created on Kaggle for their research publications (Taib and Hyndman, 2014). Kaggle’s interface allows for Data to move through The Data Life Cycle (Data One). Looking at some visualizations of the movies and tv shows on Netflix dataset, the data reveals the top contributors to Netflix (The United States being number 1, and India as number 2), top directors, and top genres (Nurdialit, 2020). The data is interesting to several audiences- including outside of the data science field.

This particular data is saved in a .CSV format, meaning a Comma-Separated List. The requirement for CSV’s on the Kaggle platform is to have a header row with human-readable names for each field. The data is visible on kaggle.com interface but can also be opened in a Microsoft excel spreadsheet. The stored data on the EXCEL spreadsheet contains 7788 entries.

The data contains: show ID- a unique ID for each show or movie, type- movie, tv show, title, director, cast, country, date added, release year, rating, duration, listed in- drama, international, ext, and description. The data is easily accessed and read by a general audience. The easy access to the metadata makes sense to the type of data it is- important to a wider audience than just data professionals and could be read and understood by the general public. The data is under creative commons License 1.0 Universal (CC0 1.0), which means there is no copyright. Anyone can copy, modify, and distribute this data without permission.

**Metadata**

The Kaggle interface displays metadata detailing update frequency, last updated, date created, current version, provenance, collection methodology, dataset owner and maintainers, and usage information. Kaggle also allows creators to add their own tags to their data set, for users to be able to find what they need under natural language tags, such as arts, entertainment, movies and tv shows.

The metadata is compiled onto one EXCEL spreadsheet- the same sheet the data is located, easily downloaded from the Kaggle interface. The excel sheet includes metadata on show id, type (movie, tv show), title, director, cast, country, date added, release year, rating, duration, listed in (drama, international, ext), and description. The metadata is column metadata, following basic dataset metadata standards but with no set schema perceivable. However, the metadata is straightforward and easy to understand for any audience or stakeholder, and it would be easy to transfer the data into Dublin core sub tags, if the author chose to. The basic elements included are- title, description, category, keywords (tags on the Kaggle interface), modification dates, licenses, and information of the creator of the dataset (Project Open Data, 2015).

There are no publications attached to the dataset, but there are “Kaggle Kernels”, where other users use the dataset for their own python environment and can combine multiple datasets for their own research purposes. Utilizing Google, Google Scholar, the only publication (non peer reviewed) found utilizing this dataset was a Medium blog post by Sriram Sharma discussing their utilization of the data set to determine the highest rating movies and shows on Netflix (Sharma, 2020). There are a few other medium articles discussing Exploratory Data Analysis (EDA) using python and Sharma’s dataset from Kaggle. There does not seem to be any peer reviewed articles utilizing this dataset, however the dataset is fairly new, and often peer reviewed articles can take a long time to reach publication.

Overall, the free use of datasets on Kaggle, the frequent collaborations with other data scientists, and the constant update of the Netflix dataset provides interesting information to stakeholders and shows the value of open access data. The presentation of the data allows a wide community of users to view, manipulate, add to, and in general, use the data in a productive way.

**Repository**

The University of California at Irvine (UCI) Machine Learning Repository is a popular resource for machine learning data sets. The repository works to address issues with computer algorithms that deal with digital data, and use the data to solve current, 21st century problems (UCI, 2015). Research at the center focuses on text mining, information retrieval, video analysis, and many other topics relating to machine learning. Therefore, this repository was chosen to potentially host the Netflix Movies and TV shows dataset. The data file for this project- relating to Movies and TV shows could be used towards AI analysis of actors, directors, identifying similar content, understanding what’s available in various countries, and many other uses. The repository hosts datasets from a wide range of interdisciplinary fields, from life sciences, to games, to business. The Netflix dataset would probably fall under the “other” category or recommended to the repository to create a “recreation” area for users to find the datasets they need.

The process of donating a dataset to the UCI repository is made simple, as their website contains a webform, where users can upload their dataset directly into the repository. The data types listed on the webform include Multivariate, Univariate, Sequential, Time-Series, Text, and Domain-Theory. There is no file format specification, and only ask for any data related to machine learning and proven useful to the community. They also ask the task (Classification, Regression, Clustering, Casual Discovery), attribute type (categorical, integer, real), area (Life Sciences, CS/ Engineering, Social Sciences, Business, Game, Other), format type (Matrix, Non-matrix), and if the dataset contains any missing values. Overall, it seems the repository accepts a wide variety of data- as long as it relates to the field of machine learning.

Anyone can donate their data to the repository, not just those apart of the UC Irvine community. There is also an email address for the institution listed to donate data anonymously, but the process of uploading data overall does not seem to have much guidance, or human assistance for the submitter. The format requirements for data submission is minimal. Minimal documentation, but they do recommend for donors to provide the best documentation possible for the machine learning community They also recommend a standard format for data that should use to the best judgement of the creator.

Downloading data is even simpler than the submission process. The interface allows for easy navigation of the 559 total datasets currently available in the repository. Users can find the data they are interested in, and have the option of downloading the data folder, the data set description, or both. The set description contains metadata information. There are no logins or authentication steps required for download.

All of the repository’s DIPs contain a parent directory, a file for data, and a file for names. Many contain an index as well. Here is an example of one of the data folders, titled “Breast Cancer Wisconsin”.

**Index of /ml/machine-learning-databases/breast-cancer-wisconsin**

Parent Directory

Index

breast-cancer-wisconsin.data

breast-cancer-wisconsin.names

unformatted-data

wdbc.data

wdbc.names

wpbc.data

wpbc.names

Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips SVN/1.7.14 Phusion\_Passenger/4.0.53 mod\_perl/2.0.11 Perl/v5.16.3 Server at archive.ics.uci.edu Port 80

Not all DIP’s look the same due to the very loose requirements needed to submit data.

The metadata on the UCI interface contains AACR2 related content-independent information- such as author, file size, date donated, number of web hits, number of attributes and instances, associated tasks, and data set characteristics. The metadata describes the physical attributes of the data, rather than the content. The metadata seems to be generated from the multiple-choice questions asked on the dataset submission form. Each data set contains varying levels of description, as the requirements to upload are very minimal. For examples, some datasets will reveal the source, additional dataset information, attribute information, relevant papers, and papers that cite this dataset. Others are bare bones, and only have brief information with a download link to the actual data. Therefore, the metadata across all of the datasets are inconsistent. The metadata does not record post-donation usage- showing the cataloguing process is mainly in the hands of the donators.

Overall, despite the repository’s lack of uniform, UCI is a well-used repository, and the content of the Netflix dataset would suit the needs of the machine learning userbase. The metadata on Kaggle, which is now holding the dataset, reflects the minimal needs required for the UCI repository. The interface helps users see who has cited or used the data for their papers. Although UCI requires a lot of upkeep and cataloguing efforts from the content creator, the repository would be a solid host for the data to be found and used by researchers interesting in Movie and TV data.

**Copyright**

The original dataset is copyrighted as CC0: Public domain, and the data should remain public domain/open access (Creative Commons, 2019). Therefore, the data should be publicly accessible in a non-proprietary format. A non-proprietary format will also aid in future, long term preservation efforts.

When data is used, a recommended citation, in following Datacite format, would appear like this:

Bansal, Shivam (2019) Netflix Movies and TV Shows. University of California, Irvine Machine Learning Repository; (URL).

**Preservation**

As mentioned before, the data will be accessible, and stored in a non-proprietary format. Therefore, the data will not be limited to specific software for users to open any files, and the data can be easily maintained by the repository hosts/data creator.

**References**

Bansal, S. (2019). Netflix Movies and TV Shows (Version 4). [Data Set]. <https://www.kaggle.com/shivamb/netflix-shows>

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml

Federal Enterprise Data Resources (2015). Project Open Data. <https://project-open-data.cio.gov/v1.1/schema/>

Nurdialit, D. (2020, September 27). Netflix movies and tv shows- exploratory data analysis (EDA) visualization using Python. *Medium. https://medium.com/analytics-vidhya/netflix-movies-and-tvshows-exploratory-data-analysis-eda-and-visualization-using-python-80753fcfcf7*

Sharma, S. (2020, February 15). A data-driven look at what’s worth watching on Netflix. *Medium.* <https://svsharma.medium.com/a-data-driven-look-at-whats-worth-watching-on-netflix-7f5e13db027e>

Taib, B. Hyndman, R.J. (2014). A gradient boosting approach to the Kaggle forecasting competition. *International Journal of Forecasting, 30*(2), 382-394.

<https://www.kaggle.com/shivamb/netflix-shows>