# **Summary**

This dataset describes 5-star rating activity from [MovieLens](http://movielens.org), a movie recommendation service. It contains 4899834 ratings across 24045 movies. These data were created by 25789 users until Nov 7th, 2016. Only users with valid five digit US postal code and having at least 20 ratings were selected. Postal code information is included for each user in the dataset.

The data is part of the paper published in LocalRec’15 workshop @RecSys 2015. (Note the data contains more recent ratings by users over the past year of 2016 but the users are the same.)

Vikas Kumar, Daniel Jarratt, Rahul Anand, Joseph A. Konstan, Brent Hecht. 2016. *"Where Far Can Be Close": Finding Distant Neighbors In Recommendation Systems*. LocalRec, ACM RecSys 1st workshop on Location Aware Recommendations 2015| [Paper](http://ceur-ws.org/Vol-1405/paper-03.pdf)| [Slides](http://www.slideshare.net/VikasKumar528/where-far-can-be-close-finding-distant-neighbors-in-recommendation-systems)

It contains three files: user\_zip.dat, movies.dat and ratings.dat

This can be downloaded from: GDrive [link](https://drive.google.com/drive/folders/0B5KFuaY1wpXoU0ZBZHU1eU1ocFU?usp=sharing)

Other GroupLens data sets are publicly available for download at<http://grouplens.org/datasets/>.

# **Usage License**

Neither the University of Minnesota nor any of the researchers involved can guarantee the correctness of the data, its suitability for any particular purpose, or the validity of results based on the use of the data set. The data set may be used for any research purposes under the following conditions:

* The user may not state or imply any endorsement from the University of Minnesota or the GroupLens Research Group.
* The user must acknowledge the use of the data set in publications resulting from the use of the data set (see below for citation information).
* The user may not redistribute the data without separate permission.
* The user may not use this information for any commercial or revenue-bearing purposes without first obtaining permission from a faculty member of the GroupLens Research Project at the University of Minnesota.
* The executable software scripts are provided "as is" without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of them is with you. Should the program prove defective, you assume the cost of all necessary servicing, repair or correction.

In no event shall the University of Minnesota, its affiliates or employees be liable to you for any damages arising out of the use or inability to use these programs (including but not limited to loss of data or data being rendered inaccurate).

If you have any further questions or comments, please email grouplens-info@umn.edu

# **Citation**

To acknowledge use of the dataset in publications, please cite the following papers:

1. Vikas Kumar, Daniel Jarratt, Rahul Anand, Joseph A. Konstan, Brent Hecht. 2016. *"Where Far Can Be Close": Finding Distant Neighbors In Recommendation Systems*. LocalRec, ACM RecSys 1st workshop on Location Aware Recommendations 2015. <http://ceur-ws.org/Vol-1405/paper-03.pdf>
2. F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=<http://dx.doi.org/10.1145/2827872>

# **Further Information About GroupLens**

GroupLens is a research group in the Department of Computer Science and Engineering at the University of Minnesota. Since its inception in 1992, GroupLens's research projects have explored a variety of fields including:

* recommender systems
* online communities
* mobile and ubiquitious technologies
* digital libraries
* local geographic information systems

GroupLens Research operates a movie recommender based on collaborative filtering, MovieLens, which is the source of these data. We encourage you to visit<http://movielens.org> to try it out! If you have exciting ideas for experimental work to conduct on MovieLens, send us an email at grouplens-info@cs.umn.edu - we are always interested in working with external collaborators.

# **Content and Use of Files**

## **Formatting and Encoding**

The dataset files are written as [“::” (two colons)](http://en.wikipedia.org/wiki/Comma-separated_values) separated files with a single header row. These files are encoded as UTF-8. If accented characters in movie titles (e.g. MisÃ©rables, Les (1995)) display incorrectly, make sure that any program reading the data, such as a text editor, terminal, or script, is configured for UTF-8.

## **User Ids**

The MovieLens users who provided valid five digit US postal code and have at least rated 20 movies were selected for inclusion. Their ids have been anonymized. These ids are consistent across the rating and postal code data.

## **Movie Ids**

Only movies with at least one rating or tag are included in the dataset. These movie ids are consistent with those used on the MovieLens web site (e.g., id 1 corresponds to the URL<https://movielens.org/movies/1>). Movie ids are consistent between ratings.dat, movies.dat.

## **Ratings Data File Structure (ratings.dat)**

All ratings are contained in the file ratings.dat. Each line of this file after the header row represents one rating of one movie by one user, and has the following format:

userId,movieId,rating,timestamp (in unix format)

The lines within this file are ordered first by userId, then, within user, by movieId.

Ratings are made on a 5-star scale, with half-star increments (0.0 stars - 5.0 stars).

Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

## **Movies Data File Structure (movies.dat)**

Movie information is contained in the file movies.csv. Each line of this file after the header row represents one movie, and has the following format:

movieId,title,genres

Movie titles are entered manually or imported from<https://www.themoviedb.org/>, and include the year of release in parentheses. Errors and inconsistencies may exist in these titles.

Genres are a pipe-separated list, and are selected from the following:

* Action
* Adventure
* Animation
* Children's
* Comedy
* Crime
* Documentary
* Drama
* Fantasy
* Film-Noir
* Horror
* Musical
* Mystery
* Romance
* Sci-Fi
* Thriller
* War
* Western
* (no genres listed)