Summary

This data set contains 10000054 ratings and 95580 tags applied to 10681 movies by 71567 users of the online movie recommender service MovieLens.

Users were selected at random for inclusion. All users selected had rated at least 20 movies. Unlike previous MovieLens data sets, no demographic information is included. Each user is represented by an id, and no other information is provided.

The data are contained in three files, `movies.dat`, `ratings.dat` and `tags.dat`. Also included are scripts for generating subsets of the data to support five-fold cross-validation of rating predictions. More details about the contents and use of all these files follows.

This and other GroupLens data sets are publicly available for download at GroupLens Data Sets.

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

Citation

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

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

Acknowledgements

Thanks to Rich Davies for generating the data set.

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' research projects have explored a variety of fields including:

- Information Filtering
- 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.

Content and Use of Files

Character Encoding

The three data files are encoded as UTF-8. This is a departure from previous MovieLens data sets, which used different character encodings. If accented characters in movie titles or tag values (e.g. Misérables, 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

Movielens users were selected at random for inclusion. Their ids have been anonymized.

Users were selected separately for inclusion in the ratings and tags data sets, which implies that user ids may appear in one set but not the other.

The anonymized values are consistent between the ratings and tags data files. That is, user id *n*, if it appears in both files, refers to the same real MovieLens user.

Ratings Data File Structure

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

```
UserID::MovieID::Rating::Timestamp
```

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.

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

Tags Data File Structure

All tags are contained in the file `tags.dat`. Each line of this file represents one tag applied to one movie by one user, and has the following format:

```
UserID::MovieID::Tag::Timestamp
```

The lines within this file are ordered first by UserID, then, within user, by MovieID.

Tags are user generated metadata about movies. Each tag is typically a single word, or short phrase. The meaning, value and purpose of a particular tag is determined by each user.

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

Movies Data File Structure

Movie information is contained in the file `movies.dat`. Each line of this file represents one movie, and has the following format:

```
MovieID::Title::Genres
```

MovieID is the real MovieLens id.

Movie titles, by policy, should be entered identically to those found in IMDB, including year of release. However, they are entered manually, so errors and inconsistencies may exist.

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

Cross-Validation Subset Generation Scripts

A Unix shell script, `split_ratings.sh`, is provided that, if desired, can be used to split the ratings data for five-fold cross-validation of rating predictions. It depends on a second script, allbut.pl, which is also included and is written in Perl. They should run without modification under Linux, Mac OS X, Cygwin or other Unix like systems.

Running `split_ratings.sh` will use `ratings.dat` as input, and produce the fourteen output files described below. Multiple runs of the script will produce identical results.

File Names	Description
r1.train, r2.train, r3.train, r4.train, r5.train r1.test, r2.test, r3.test, r4.test, r5.test	The data sets r1.train and r1.test through r5.train and r5.test are 80%/20% splits of the ratings data into training and test data. Each of r1, ..., r5 have disjoint test sets; this if for 5 fold cross validation (where you repeat your experiment with each training and test set and average the results).
ra.train, rb.train ra.test, rb.test	The data sets ra.train, ra.test, rb.train, and rb.test split the ratings data into a training set and a test set with exactly 10 ratings per user in the test set. The sets ra.test and rb.test are disjoint.