Music lastfm-2k/readme.md

===================

hetrec2011-lastfm-2k

===================

Version

Version 1.0 (May 2011)

Description

This dataset contains social networking, tagging, and music artist listening information
from a set of 2K users from Last.fm online music system.
http://www.last.fm

The dataset is released in the framework of the 2nd International Workshop on
Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011)
http://ir.ii.uam.es/hetrec2011
at the 5th ACM Conference on Recommender Systems (RecSys 2011)
http://recsys.acm.org/2011

Data statistics

1892 users
17632 artists

12717 bi-directional user friend relations, i.e. 25434 (user_i, user_j) pairs
avg. 13.443 friend relations per user

92834 user-listened artist relations, i.e. tuples [user, artist, listeningCount]
avg. 49.067 artists most listened by each user
avg. 5.265 users who listened each artist

11946 tags

186479 tag assignments (tas), i.e. tuples [user, tag, artist]

avg. 98.562 tas per user

avg. 14.891 tas per artist

avg. 18.930 distinct tags used by each user

avg. 8.764 distinct tags used for each artist

Files

-----

* artists.dat

This file contains information about music artists listened and tagged by the users.

* tags.dat

This file contains the set of tags available in the dataset.

* user_artists.dat

This file contains the artists listened by each user.

It also provides a listening count for each [user, artist] pair.

* user_taggedartists.dat - user_taggedartists-timestamps.dat

These files contain the tag assignments of artists provided by each particular user.

They also contain the timestamps when the tag assignments were done.

* user_friends.dat

These files contain the friend relations between users in the database.

Data format

The data is formatted one entry per line as follows (tab separated, "\t"):

* artists.dat

```
id \t name \t url \t pictureURL
```

Example:

```
707	Metallica	http://www.last.fm/music/Metallica	http://userserve-ak.last.fm/serve/252/7560709.jpg
```

* tags.dat

```
tagID \t tagValue
1	metal
```

* user_artists.dat

```
userID \t artistID \t weight
2	51	13883
```

* user_taggedartists.dat

```
userID \t artistID \t tagID \t day \t month \t year
2	52	13	1	4	2009
```

* user_taggedartists-timestamps.dat

```
userID \t artistID \t tagID \t timestamp
2	52	13	1238536800000
```

* user_friends.dat

```
userID \t friendID
2	275
```

License

The users' names and other personal information in Last.fm are not provided in the dataset.

The data contained in hetrec2011-lastfm-2k.zip is made available for non-commercial use.

Those interested in using the data in a commercial context should contact Last.fm staff:
http://www.lastfm.com/about/contact

Acknowledgements

This work was supported by the Spanish Ministry of Science and Innovation (TIN2008-06566-C04-02),
and the Regional Government of Madrid (S2009TIC-1542).

References

When using this dataset you should cite:

- Last.fm website, http://www.lastfm.com

You may also cite HetRec'11 workshop as follows:

```
@inproceedings{Cantador:RecSys2011,
  author = {Cantador, Iv\'{a}n and Brusilovsky, Peter and Kuflik, Tsvi},
  title = {2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011)},
  booktitle = {Proceedings of the 5th ACM conference on Recommender systems},
  series = {RecSys 2011},
  year = {2011},
  location = {Chicago, IL, USA},
  publisher = {ACM},
  address = {New York, NY, USA},
  keywords = {information heterogeneity, information integration, recommender systems},
}
```

Credits

This dataset was built by Ignacio Fern�ndez-Tob�as with the collaboration of Iv�n Cantador and Alejandro Bellog�n,

members of the Information Retrieval group at Universidad Autonoma de Madrid (http://ir.ii.uam.es)

Contact

Iv�n Cantador, ivan [dot] cantador [at] uam [dot] es

Movie ml-1m/readme.md

SUMMARY

==

These files contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000.

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.

If you have any further questions or comments, please contact GroupLens <grouplens-info@cs.umn.edu>.

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 Shyong Lam and Jon Herlocker for cleaning up and generating the data set.

FURTHER INFORMATION ABOUT THE GROUPLENS RESEARCH PROJECT
==

The GroupLens Research Project is a research group in the Department of Computer Science and Engineering at the University of Minnesota. Members of the GroupLens Research Project are involved in many research projects related to the fields of information filtering, collaborative filtering, and recommender systems. The project is lead by professors John Riedl and Joseph Konstan. The project began to explore automated collaborative filtering in 1992, but is most well known for its world wide trial of an automated collaborative filtering system for Usenet news in 1996. Since then the project has expanded its scope to research overall information filtering solutions, integrating in content-based methods as well as improving current collaborative filtering technology.

Further information on the GroupLens Research project, including research publications, can be found at the following web site:

    http://www.grouplens.org/

GroupLens Research currently operates a movie recommender based on collaborative filtering:

    http://www.movielens.org/

RATINGS FILE DESCRIPTION
==

All ratings are contained in the file "ratings.dat" and are in the following format:

UserID::MovieID::Rating::Timestamp

- UserIDs range between 1 and 6040
- MovieIDs range between 1 and 3952
- Ratings are made on a 5-star scale (whole-star ratings only)
- Timestamp is represented in seconds since the epoch as returned by time(2)
- Each user has at least 20 ratings

USERS FILE DESCRIPTION
==

User information is in the file "users.dat" and is in the following format:

UserID::Gender::Age::Occupation::Zip-code

All demographic information is provided voluntarily by the users and is not checked for accuracy. Only users who have provided some demographic information are included in this data set.

- Gender is denoted by a "M" for male and "F" for female
- Age is chosen from the following ranges:

    * 1: "Under 18"
    * 18: "18-24"
    * 25: "25-34"
    * 35: "35-44"
    * 45: "45-49"
    * 50: "50-55"
    * 56: "56+"

- Occupation is chosen from the following choices:

* 0: "other" or not specified
* 1: "academic/educator"
* 2: "artist"
* 3: "clerical/admin"
* 4: "college/grad student"
* 5: "customer service"
* 6: "doctor/health care"
* 7: "executive/managerial"
* 8: "farmer"
* 9: "homemaker"
* 10: "K-12 student"
* 11: "lawyer"
* 12: "programmer"
* 13: "retired"
* 14: "sales/marketing"
* 15: "scientist"
* 16: "self-employed"
* 17: "technician/engineer"
* 18: "tradesman/craftsman"
* 19: "unemployed"
* 20: "writer"

MOVIES FILE DESCRIPTION

===

Movie information is in the file "movies.dat" and is in the following
format:

MovieID::Title::Genres

- Titles are identical to titles provided by the IMDB (including
year of release)
- Genres are pipe-separated and are selected from the following genres:

* Action
* Adventure

* Animation

* Children's

* Comedy

* Crime

* Documentary

* Drama

* Fantasy

* Film-Noir

* Horror

* Musical

* Mystery

* Romance

* Sci-Fi

* Thriller

* War

* Western

- Some MovieIDs do not correspond to a movie due to accidental duplicate entries and/or test entries
- Movies are mostly entered by hand, so errors and inconsistencies may exist

Movie ml-100k

SUMMARY & USAGE LICENSE

===

MovieLens data sets were collected by the GroupLens Research Project at the University of Minnesota.

This data set consists of:

* 100,000 ratings (1-5) from 943 users on 1682 movies.
* Each user has rated at least 20 movies.
* Simple demographic info for the users (age, gender, occupation, zip)

The data was collected through the MovieLens web site (movielens.umn.edu) during the seven-month period from September 19th, 1997 through April 22nd, 1998. This data has been cleaned up - users who had less than 20 ratings or did not have complete demographic information were removed from this data set. Detailed descriptions of the data file can be found at the end of this file.

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.

If you have any further questions or comments, please contact GroupLens <grouplens-info@cs.umn.edu>.

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 Al Borchers for cleaning up this data and writing the accompanying scripts.

PUBLISHED WORK THAT HAS USED THIS DATASET

==

Herlocker, J., Konstan, J., Borchers, A., Riedl, J.. An Algorithmic Framework for Performing Collaborative Filtering. Proceedings of the 1999 Conference on Research and Development in Information Retrieval. Aug. 1999.

FURTHER INFORMATION ABOUT THE GROUPLENS RESEARCH PROJECT

==

The GroupLens Research Project is a research group in the Department of Computer Science and Engineering at the University of Minnesota. Members of the GroupLens Research Project are involved in many research projects related to the fields of information filtering, collaborative filtering, and recommender systems. The project is lead by professors John Riedl and Joseph Konstan. The project began to explore automated collaborative filtering in 1992, but is most well known for its world wide trial of an automated collaborative filtering system for Usenet news in 1996. The technology developed in the Usenet trial formed the base for the formation of Net Perceptions, Inc., which was founded by members of GroupLens Research. Since then the project has expanded its scope to research overall information filtering solutions, integrating in content-based methods as well as improving current collaborative filtering technology.

Further information on the GroupLens Research project, including research publications, can be found at the following web site:

http://www.grouplens.org/

GroupLens Research currently operates a movie recommender based on collaborative filtering:

http://www.movielens.org/

DETAILED DESCRIPTIONS OF DATA FILES
==

Here are brief descriptions of the data.

ml-data.tar.gz   -- Compressed tar file. To rebuild the u data files do this:

```
gunzip ml-data.tar.gz
tar xvf ml-data.tar
mku.sh
```

u.data   -- The full u data set, 100000 ratings by 943 users on 1682 items.
Each user has rated at least 20 movies. Users and items are
numbered consecutively from 1. The data is randomly
ordered. This is a tab separated list of
user id | item id | rating | timestamp.
The time stamps are unix seconds since 1/1/1970 UTC

u.info -- The number of users, items, and ratings in the u data set.

u.item -- Information about the items (movies); this is a tab separated
list of
movie id | movie title | release date | video release date |
IMDb URL | unknown | Action | Adventure | Animation |
Children's | Comedy | Crime | Documentary | Drama | Fantasy |
Film-Noir | Horror | Musical | Mystery | Romance | Sci-Fi |
Thriller | War | Western |
The last 19 fields are the genres, a 1 indicates the movie
is of that genre, a 0 indicates it is not; movies can be in
several genres at once.
The movie ids are the ones used in the u.data data set.

u.genre -- A list of the genres.

u.user -- Demographic information about the users; this is a tab
separated list of
user id | age | gender | occupation | zip code
The user ids are the ones used in the u.data data set.

u.occupation -- A list of the occupations.

u1.base -- The data sets u1.base and u1.test through u5.base and u5.test
u1.test are 80%/20% splits of the u data into training and test data.
u2.base Each of u1, ..., u5 have disjoint test sets; this if for
u2.test 5 fold cross validation (where you repeat your experiment
u3.base with each training and test set and average the results).
u3.test These data sets can be generated from u.data by mku.sh.
u4.base
u4.test
u5.base
u5.test

ua.base -- The data sets ua.base, ua.test, ub.base, and ub.test
ua.test split the u data into a training set and a test set with
ub.base exactly 10 ratings per user in the test set. The sets
ub.test ua.test and ub.test are disjoint. These data sets can

be generated from u.data by mku.sh.

allbut.pl -- The script that generates training and test sets where
all but n of a users ratings are in the training data.

mku.sh -- A shell script to generate all the u data sets from u.data.