**总结**

此数据集（ml-latest-small）描述了[MovieLens]的五星评级和自由文本标记活动(http://movielens.org)，电影推荐服务。它包含9742部电影的100836个打分和3683个标签。这些数据由610名用户在1996年3月29日至2018年9月24日期间创建。该数据集于2018年9日26日生成。

用户都是随机选择的。所有选定的用户都至少对20部电影进行了评级。每个用户由一个id表示，除此之外不提供其他信息。

数据包含`link.csv`，`movies.csv`，`ratings.csv`和 `tags.csv`。以下是所有这些文件的内容和使用的更多详细信息。

这是一个\*开发\*数据集。因此，它可能会随着时间的推移而改变，不是共享研究结果的合适数据集。你也可以查看可用的\*基准\*数据集。GroupLens Research运营着一个基于协作过滤的电影推荐器MovieLens，这是这些数据的来源。我们鼓励您访问<http://movielens.org>！如果您对MovieLens的实验工作有想法请发送电子邮件至<grouplens-info@cs.umn.edu>！

**文件的内容和使用**

1. 格式化和编码

数据集文件写为[逗号分隔的值](http://en.wikipedia.org/wiki/Comma-separated\_values)具有单个标题行的文件。（就是csv）。包含逗号（`，`）的列使用双引号（`“`）进行转义。这些文件编码为UTF-8。如果电影标题或标记值（例如Misérables，Les（1995））中的重音字符显示不正确，请确保任何读取数据的程序（如文本编辑器、终端或脚本）都配置为UTF-8。

时间戳表示自1970年1月1日午夜协调世界时（UTC）以来的秒数。

1. User ID

用户ID在`ratings.csv`和 `tag.csv` 都是一样的（即，相同的id表示两个文件中的相同用户）。

1. Movies ID

数据集中包含至少一个分级或标记的电影。这些电影id与MovieLens网站上使用的id一致（例如，id“1”对应于URL<https://movielens.org/movies/1>). 在 `ratings.csv`，`tag.csv`，`movies.csv`和`link.csv`中电影ID都是一致的。

1. Ratings.csv

该文件标题行后的每一行表示一个用户对一部电影的一个分级，格式如下：

userId，movieId，rating，timestamp

此文件中的行首先按userId排序，然后在user中按movieId排序。

评级以五星制，以半星为增量（0.5星-5.0星）。

1. Tags.csv

该文件标题行后的每一行表示一个用户应用于一部电影的一个标记，格式如下：

userId，movieId，tag，timestamp

此文件中的行首先按userId排序，然后在user中按movieId排序。

标签是用户生成的关于电影的元数据。每个标签通常是单个单词或短语。特定标记的含义、值和用途由每个用户决定。

1. Movies.csv

标题行之后的该文件的每一行代表一部电影，格式如下：

movieId，title，genres

电影标题是手动输入的或从导入的<https://www.themoviedb.org/>，并在括号中包括发布年份。这些标题中可能存在错误和不一致。

genres是一个管道分隔列表，可从以下选项中选择：

\*行动\*冒险\*动画\*儿童的\*喜剧\*犯罪\*纪录片\*戏剧\*幻想\*黑色电影\*恐怖\*音乐剧\*神秘\*浪漫\*科幻小说\*颤栗\*战争\*西部\*（未列出流派）

1. Links.csv

可用于链接到其他电影数据源的标识符包含在文件“links.csv”中。标题行之后的该文件的每一行代表一部电影，格式如下：

MoviesID、imdbId、tmdbId

movieId是由使用的电影的标识符<https://movielens.org>. E、 例如，电影《玩具总动员》有链接<https://movielens.org/movies/1>.

imdbId是由使用的电影的标识符<http://www.imdb.com>. E、 例如，电影《玩具总动员》有链接<http://www.imdb.com/title/tt0114709/>.

tmdbId是使用的电影的标识符<https://www.themoviedb.org>. E、 例如，电影《玩具总动员》有链接<https://www.themoviedb.org/movie/862>.

1. 交叉验证

MovieLens数据集的早期版本包括预先计算的交叉折叠或脚本来执行此计算。我们不再将这些功能与数据集捆绑在一起，因为大多数现代工具包都将其作为内置功能提供。如果您想了解推荐系统评估中交叉折叠计算的标准方法，请参见[LensKit](http://lenskit.org)用于工具、文档和开源代码示例。