# 摘要

该数据集（ml-latest）描述了电影推荐服务[MovieLens](http://movielens.org/)的五星级评级和自由文本标记活动。它包含58098部电影中的27753444个评分和1108997个标签应用程序。这些数据由1995年1月9日至2018年9月26日之间的283228用户创建。此数据集于2018年9月26日生成。

随机选择要纳入的用户。所有选定的用户都评价了至少1部电影。不包括人口统计信息。每个用户用一个id表示，并且不提供其他信息。

该数据包含在文件中genome-scores.csv，genome-tags.csv，links.csv，movies.csv，ratings.csv和tags.csv。以下是有关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，它可能会随着时间而改变，而不是共享研究结果的合适数据集。如果您打算这样做，请查看可用的基准数据集。

此和其他GroupLens数据集可从<http://grouplens.org/datasets/>公开下载。

# 使用许可

明尼苏达大学或所涉研究人员均不能保证数据的正确性，对任何特定目的的适用性或基于使用数据集的结果的有效性。在以下情况下，该数据集可用于任何研究目的：

* 用户不得声明或暗示明尼苏达大学或GroupLens研究小组的任何认可。
* 用户必须确认由于使用数据集而在出版物中使用了该数据集（有关引用信息，请参见下文）。
* 未经单独许可，用户不得重新分发数据。
* 未经明尼苏达大学的GroupLens研究项目的教职人员首先获得许可，用户不得将这些信息用于任何商业或收益目的。
* 可执行软件脚本按“原样”提供，没有任何形式的明示或暗示保证，包括但不限于针对特定目的的适销性和适用性的暗示保证。它们的质量和性能的全部风险都由您承担。如果程序证明有缺陷，则您承担所有必要的维修，修理或校正费用。

对于因使用或无法使用这些程序而引起的任何损失（包括但不限于数据丢失或数据不准确），明尼苏达大学，其附属机构或员工绝不对您承担任何责任。

如果您还有其他问题或意见，请发送电子邮件[至grouplens-info@umn.edu](mailto:grouplens-info@umn.edu)

# 引文

要确认出版物中使用了数据集，请引用以下论文：

F. Maxwell Harper和Joseph A. Konstan。2015年。MovieLens数据集：历史和上下文。交互式智能系统（TiiS）5、4：19：1–19：19上的ACM事务。<https://doi.org/10.1145/2827872>

# 有关GroupLens的更多信息

GroupLens是明尼苏达大学计算机科学与工程系的一个研究小组。自1992年成立以来，GroupLens的研究项目已经探索了多个领域，包括：

* 推荐系统
* 在线社区
* 移动和无处不在的技术
* 数字图书馆
* 当地地理信息系统

GroupLens Research基于协作过滤MovieLens来操作电影推荐器，MovieLens是这些数据的来源。我们鼓励您访问[http://movielens.org](http://movielens.org/)进行尝试！如果您对在MovieLens上进行实验工作有令人振奋的想法，请给我们发送电子邮件至[grouplens-info@cs.umn.edu-](mailto:grouplens-info@cs.umn.edu)我们始终对与外部合作者合作感兴趣。

# 文件内容和使用

## 格式化和编码

数据集文件被编写为带有单个标题行的[逗号分隔值](http://en.wikipedia.org/wiki/Comma-separated_values)文件。包含逗号（,）的列使用双引号（"）进行转义。这些文件被编码为UTF-8。如果电影标题或标签值中的重音字符显示不正确（例如Misérables，Les（1995）），请确保为UTF-8配置了任何读取数据的程序，例如文本编辑器，终端或脚本。

## 用户编号

随机选择了MovieLens用户进行收录。他们的身份证已经匿名。ratings.csv和之间的用户ID保持一致tags.csv（即，相同的ID指的是两个文件中的同一用户）。

## 电影ID

数据集中仅包含具有至少一个等级或标签的电影。这些电影ID与MovieLens网站上使用的电影ID一致（例如，ID 1对应于URL <https://movielens.org/movies/1>）。电影ID是间保持一致ratings.csv，tags.csv，movies.csv，和links.csv（即相同的ID是指横跨这四个数据文件相同的电影）。

## 评分数据文件结构（ratings.csv）

所有评级都包含在文件中ratings.csv。标题行之后的此文件的每一行代表一个用户对一部电影的评分，并具有以下格式：

userId,movieId,rating,timestamp

该文件中的各行首先由userId排序，然后在用户中由movieId排序。

评分以5星级标准进行，以半星级递增（0.5星级-5.0星级）。

时间戳表示自1970年1月1日午夜协调世界时（UTC）起的秒数。

## 标签数据文件结构（tags.csv）

所有标签都包含在文件中tags.csv。标题行之后的该文件的每一行代表一个用户将一个标签应用于一部电影，并具有以下格式：

userId,movieId,tag,timestamp

该文件中的各行首先由userId排序，然后在用户中由movieId排序。

标签是用户生成的有关电影的元数据。每个标签通常是一个单词或简短短语。特定标签的含义，值和目的由每个用户确定。

时间戳表示自1970年1月1日午夜协调世界时（UTC）起的秒数。

## 电影数据文件结构（movies.csv）

电影信息包含在文件中movies.csv。标题行之后的此文件的每一行代表一部电影，并具有以下格式：

movieId,title,genres

影片标题可手动输入或从<https://www.themoviedb.org/>导入，并在括号中包含发行年份。这些标题中可能存在错误和不一致。

流派是用管道分隔的列表，并从以下列表中选择：

* 行动
* 冒险
* 动画
* 儿童的
* 喜剧
* 犯罪
* 记录
* 戏剧
* 幻想
* 黑色电影
* 恐怖
* 音乐
* 神秘
* 浪漫
* 科幻
* 惊悚片
* 战争
* 西
* （未列出流派）

## 链接数据文件结构（links.csv）

文件中包含可用于链接到其他电影数据源的标识符links.csv。标题行之后的此文件的每一行代表一部电影，并具有以下格式：

movieId,imdbId,tmdbId

movieId是[https://movielens.org](https://movielens.org/)使用的电影的标识符。例如，电影《玩具总动员》具有链接<https://movielens.org/movies/1>。

imdbId是[http://www.imdb.com](http://www.imdb.com/)使用的电影的标识符。例如，电影《玩具总动员》具有链接<http://www.imdb.com/title/tt0114709/>。

tmdbId是[https://www.themoviedb.org](https://www.themoviedb.org/)使用的电影的标识符。例如，电影《玩具总动员》具有链接<https://www.themoviedb.org/movie/862>。

上面所列资源的使用受每个提供商的条款约束。

## 标记基因组（genome-scores.csv和基因组标签.csv）

该数据集包括标签基因组的当前副本。

标签基因组是一种数据结构，其中包含电影的标签相关性得分。该结构是一个密集的矩阵：基因组中的每部电影都有一个值，用于基因组中的每个标签。

如[本文](http://files.grouplens.org/papers/tag_genome.pdf)所述，标签基因组编码电影以多强的强度表现出标签所代表的特定属性（大气，发人深省，逼真的等）。标签基因组是使用机器学习算法对用户提供的内容（包括标签，评分和文字评论）进行计算的。

基因组被分成两个文件。该文件genome-scores.csv包含以下格式的电影标签相关性数据：

movieId,tagId,relevance

第二个文件，genome-tags.csv以以下格式提供了基因组文件中标签ID的标签描述：

tagId,tag

这些tagId值是在导出数据集时生成的，因此，每个MovieLens数据集的版本可能会有所不同。

如果引用标签基因组数据，请包括以下引用：

Jesse Vig，Shilad Sen和John Riedl。2012年。标签基因组：编码社区知识以支持新型交互。ACM Trans。相互作用。智力 Syst。2：3：13：1–13：44。<https://doi.org/10.1145/2362394.2362395>

## 交叉验证

MovieLens数据集的先前版本包含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中的任何一个与数据集捆绑在一起，因为大多数现代工具包都将此功能作为内置功能提供。如果您希望在推荐系统评估的背景下了解交叉计算的标准方法，请参阅[LensKit](http://lenskit.org/)以获取工具，文档和开源代码示例。