数据集使用说明

1 权利声明

本数据集抓取自 Lastfm，所有数据归 Lastfm 所有，禁止商业用途。

如果您想使用该数据集进行科研活动，请务必给出对 Last.fm 及本文的引用信息。

2 数据特点

1. 包含完整的用户、歌曲、曲作者的基本信息。

2. 包含丰富的用户行为记录，可用于构造用户行为序列。

3. 包含歌曲、曲作者的显著标签信息，可用于从文本的角度描述歌曲和作者。

4. 提供了标签的基本信息。

5. 数据被随机分组，可直接用来实验。

3 组织形式

本数据集使用 Mysql 进行管理，对应的数据库名为 lastfm，您可以非常容易地将其导入并使用。数据集包含有 5 个基本的数据表：

记录表 record 用于记录用户的收听行为，如某用户在某时间段收听了某歌曲。record 由记录标识符(rid:int)、用户标识符(uid:varchar)、歌曲 mbid(mbid:varchar)、记录发生的 unix 时间戳(uts:varchar)、记录发生的日期时间(datetime:varchar)、记录所属分组(scale:int)等字段构成。

用户表 user 用于记录用户的基本信息，其由用户标识符(uid:varchar)、用户名(username:varchar)、用户国籍(country:varchar)、用户年龄(age:varchar)、用户性别(gender:varchar)、用户注册时的 unix 时间戳(registeredTime:varchar)、用户注册日期时间(registeredText:varchar)、播放序列(playlist:text)、用户所属分组(scale:int)等字段构成。

歌曲表 song 用于记录歌曲的基本信息，其由歌曲标识符(sid:varchar)、歌曲对应 mbid（mbid:varchar）、歌曲名称(name:text)、歌曲时长(duration:varchar)、曲作者标识符(aid:varchar)、曲作者名称(aname:varchar)、专辑名(album:text)、听众数目(listeners:varchar)、播放次数(playcount:varchar)、描述歌曲的显著标签

(toptag:text)等字段构成。

曲作者表 artist 用于记录曲作者的基本信息，其由曲作者标识符(mbid:varchar)、曲作者名称(name:text)、曲作者图片的链接(img:text)和描述曲作者的显著标签(toptag:text)构成。

标签表 tag 用于记录标签的基本信息，其由标签标识符(id:varchar)、标签名称（name:text）、标签被创建的次数(reach:varchar)、标签被使用的次数(taggings:text)等字段构成。

4 字段解析

4.1 scale

记录表 record 和用户 user 中的 scale 字段用以表征记录和用户所处的分组编号。为了方便，本数据集将用户记录和用户分为 Unused、Small、Whole 和 Session 四类。其中，Small、Whole 和 Session 被 scale 字段分割成 40 组，其中第 0 组到第 9 组属于 Small 数据集，第 0 组到第 29 组属于 Whole 数据集，第 30 组到第 39 组属于 Session 数据集。Unused 数据的 scale 设为-1。显然，Small 数据集是 Whole 数据集的一部分，它们的特点是**每一个用户所收听的歌曲都在一个会话期内，即不存在长时中断**。从 Whole 数据集中划分出 Small 数据集的主要目的是方便机器性能不佳的用户使用，对于 Small 数据集，用户可以使用 10 组中的 9 组作训练集而余下的一组作测试集。Session 数据集与 Whole 数据集的主要区别是**每一个用户所收听的歌曲至少在两个会话期内**。类似的，用户可以用其中 9 组作训练集而余下的一组作测试集。下表给出了 Small、Whole 和 Session 三类数据集的基本统计信息。

表 1. 不同数据集的统计信息

	Small	Whole	Session
用户数	1530	4590	1690
歌曲数	24992	62422	32218
稀疏度	99.92%	99.97%	99.92%
最大长度	30	30	66
最短长度	10	10	20
中位长度	24	24	30

4.2 playlist

数据表 user 中的 playlist 字段用以表征用户按序收听的歌曲构成的序列，数据如“sid1:ratio1==>sid2:ratio2==>...==>sidn:ration”所示。其中，sid 表示被听歌曲的标识符(注:非 mbid)。ratio 表示两首歌之间的时间间隔与前一首歌曲时长的比例，用以表征用户收听该首歌曲的时长比例。显然，ratio 过小表示用户刚开始收听遍跳过，ratio 过大表明歌曲被完整收听而且还可能有暂停发生。

4.3 toptag

数据表 song 和 artist 中的字段 toptag 表示 Lastfm 网站中的用户给歌曲或曲作者所打的显著标签，数据如“{tag1:count1,tag2:count2,...,tagn:countn}”所示。其中，tag 表示被打标签名称，count 表示标签被标记次数。需要注意的是，在 Lastfm 中，count 并非标签被应用于歌曲或曲作者的绝对次数，而是标签相对于被使用最多次的标签的相对次数。例如在描述歌曲“Collapse of History”的标签中，“industrial”被使用最多次且次数为 200，而标签“Stars”被使用 100 次。那么，在歌曲记录对应的字段 toptag 中，“industrial”对应的 count 为 100，“Stars”对应的 count 为 50，即{“industrial”:100,”starts”:50}，以此类推。

4.4 其他字段

数据表中的其他字段都比较简单直观，这里就不再一一介绍。

4 应用场景

1. 使用文本分析的方法描述歌曲或者曲作者特征。
2. 分析用户所收听歌曲的序列，包括跨会话分析和会话内分析。
3. 预测用户下一首可能收听的歌曲或者曲作者。
4. 生成用户可能喜欢的播放列表。
5. 标签预测问题。
6. 其他适合的应用场景。