第五周作业 歌曲推荐系统

一、数据：Million Song Dataset(MSD)

### <https://labrosa.ee.columbia.edu/millionsong/>

作业使用的数据集是公开音乐数据集 Million Song Dataset(MSD) ， 它 包 含 来 自 SecondHandSongs dataset 、 musiXmatch dataset、Last.fm dataset、Taste Profile subset、 thisismyjam-to-MSD mapping、tagtraum genre annotations 和 Top MAGD dataset 七个知名音乐社区的数据。

原始数据集包括：

1. train\_triplets.txt：三元组数据（用户、歌曲、播放次数）
2. track\_metadata.db：每个歌曲的元数据

由于原始数据太大，作业用的数据集只是其中的子集（播放次数最多的800个用户、播放次数最多的800首歌曲。

数据预处理过程请见1.DataProcessing.ipynb文件（不必运行该程序，运行该程序需要原始数据。可以通过看代码理解数据提取过程），最后得到的数据文件为：triplet\_dataset\_sub.csv（37000+条记录）

二、作业要求：将triplet\_dataset\_sub.csv中的数据用train\_test\_split分成80%数据做训练，剩下20%数据做测试。

1. 实现基于用户的协同过滤； （20分）
2. 实现基于物品的协同过滤； （20分）
3. 实现基于模型（矩阵分解）的协同过滤。（30分）
4. 对每种推荐算法的推荐结果，用Top10个推荐歌曲的准确率和召回率评价推荐系统的性能。（30分）

注意：

1. 由于这个数据集中并没有用户对物品的显式打分，需要将播放次数转换为分数。
2. 在协同过滤中，计算用户之间的相似度或物品之间的相似度时，一种方式用播放次数/比例作为用户/物品的特征表示，同课件。

另一种可选的表示是只要用户播放过歌曲就表示为1，否则为0（二值化），这样物品之间的相似度为播放两个歌曲的用户交集除以播放两个歌曲的用户并集：

。

类似的，两个用户之间的相似度可用两个用户播放歌曲的交集除以两个用户播放歌曲的并集表示。