电影推荐系统

基本介绍

- 本项目主要包括ETL，模型选择，网站设计三个部分
- 数据集使用老师发布的采集的自豆瓣的数据集
- 使用FM模型预测评分的方式推荐电影，该模型由spark提供接口
- 网站前端使用Python的flask框架实现，后端使用hive数据库，网站通过pyspark访问数据库
- 浏览者使用会自动以uid为0的用户的身份登录，也可以切换成其他用户身份登录，本项目不考虑冷启动问题
- 浏览者可以进行的操作有切换用户和打分，推荐电影会显示在网页右侧，每次评分后会重新训练模型
- 网站当前运行在121.199.6.189:12345，因为权限原因公网无法打开，需要建立本地端口转发才能访问 。运行 `ssh -NL 12345:127.0.0.1:12345 sxy_ml6@121.199.6.189` ，然后通过http://127.0.0.1:12345访问
- 代码在https://github.com/cleanerleon/mtimes.git
- 网站部分文字和图片来自真实的http://www.mtime.com/

ETL

- 老师提供的数据集包括movie.csv和user.csv两个文件
- moive文件格式为 `类型,主演,地区,导演,特色,评分,电影名` ，包含93160条记录
- movie文件中包括一些冗余信息，如

```
剧情,徐峥|王传君|周一围|谭卓|章宇,中国大陆,文牧野,经典,8.9,我不是药神
剧情,徐峥|王传君|周一围|谭卓|章宇,中国大陆,文牧野,感人,8.9,我不是药神
喜剧,徐峥|王传君|周一围|谭卓|章宇,中国大陆,文牧野,经典,9.0,我不是药神
喜剧,徐峥|王传君|周一围|谭卓|章宇,中国大陆,文牧野,搞笑,9.0,我不是药神
喜剧,徐峥|王传君|周一围|谭卓|章宇,中国大陆,文牧野,感人,9.0,我不是药神
犯罪,徐峥|王传君|周一围|谭卓|章宇,中国大陆,文牧野,感人,9.0,我不是药神
犯罪,徐峥|王传君|周一围|谭卓|章宇,中国大陆,文牧野,搞笑,9.0,我不是药神
```

- user文件格式为 `评分,用户名,评论时间,用户ID,电影名,类型` ，包含199813条信息
- 定义如下表格，其中数据格式中包含一些冗余信息，如User、Actor、Director中都包含点评、参演、导演的电影id的列表，以方便检索

```
Movie: id bigint, name string, genres array<bigint>, actors array<bigint>, district bigint,
directors
Rating: uid bigint, mid bigint, time timestamp, rating bigint
User: id bigint, name string, movies array<bigint>array<bigint>, traits array<bigint>,
rating double
Actor: id bigint, name string, movies array<bigint>
Director: id bigint, name string, movies array<bigint>
Genre: id bigint, name string, movies array<bigint>
District: id bigint, name string, movies array<bigint>
Trait: id bigint, name string, movies array<bigint>
```

- 数据集共包括23034个部电影，14405个导演，39283个演员，13532个用户，199813条评分记录

模型选择

MF

- spark直接提供了MF的接口（pyspark.ml.recommendation.ALS），主要有3个超参数rank，maxIter和reg，支持将训练好的模型保存成文件和从文件中加载模型
- 根据测试，rank=95，maxIter=20，reg=0.6，得到最小RMSE 2.664016，但是考虑到耗时问题选择了rank和maxIter较小的方案，rank=35，maxIter=10，reg=0.8，此时RMSE为2.6774639553502992，不过模型训练时间从25秒降低到5秒
- ALS不支持冷启动，其参数coldStartStrategy设置为drop之后才能调用RegressionEvaluator评估RMSE，测试数据集如果有冷启动的样本RMSE效果特别好，可以达到1.4，可能是ALS直接讲冷启动样本的预测值设为原始值，因此在评估多个模型时需要将测试数据集中冷启动样本去掉

FM

- 使用xlearn提供的接口，主要有三个超参数k（就是rank），epoch，lambda，同样可以将训练好的模型保存成文件或者从文件中加载模型。
- 对于用户电影评分矩阵，k=64，lambda=0.001，epoch=10时 rmse最小为2.635498
- 对于将电影信息做one-hot编码导入评分矩阵，最小rmse也在2.6的水平

结论

- 考虑到FM与MF相比并无太显著的差异，同时整个系统基础设施都架构在spark上，最后选择使用spark的ALS接口

推荐流程

召回

- 选择该用户评分>6的电影，再选择这些电影的导演和演员参与的评分>6的电影，去重和去除已经评分过的电影，用作召回数据集
- 如果数据集不到50个，就选择评分最高的电影补齐

推荐

- 根据预先训练好的模型预测召回数据集的评分，排序后输出

代码组织

----data ：csv文件

|---docs：文档

|---scripts：etl代码，模型测试代码

|---server：服务器代码

部署&运行

- 将server目录拷贝到121.199.6.189
- 依赖flask，pandas，pyspark==2.3.1
- 进入server目录运行 `PYSPARK_PYTHON=python3 python3 main.py`

不足

- FM应该是更具有优势的模型，因为时间有限没有优化
- 电影推荐排序应该根据浏览页面的变化而变化，本项目为了观察用户评分对排序的影响而没有考虑其他因素仅考虑预测值
- 对hive不熟悉一些SQL可能有优化的地方，特别是hive只能增加不能删除和修改，导致一些查询比较复杂
- 在用户对电影评分后，并没有更新电影的总评分

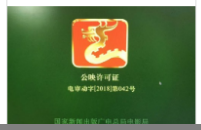
截图

时光网

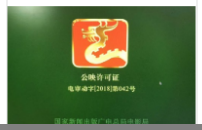
欢迎您 身似 首页 电影列表 导演列表 演员列表 切换用户

身似

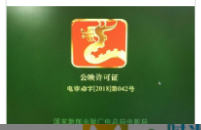
垫底辣妹
评分 9
2019-05-01 20:05



大卫和卡玛尔
评分 10
2018-01-28 14:01



快手枪手快枪手
评分 10
2018-01-28 14:01



猫咪后院之家
评分 10
2018-01-28 13:01

假装有图片

推荐电影

爱的捆绑
评分 7.3

假装有图片

少年手指虎
评分 7.2

时光网

欢迎您 身似 首页 电影列表 导演列表

请评分 ×

★ ★ ★ ★ ★ ★ ★ ★ ★ ★ 确定

垫底辣妹

总评分 8.1 打分

- 导演： 土井裕泰
- 主演： 有村架纯 伊藤淳史 吉田羊 田中哲司 野村周平
- 类型： 剧情 喜剧
- 地区： 日本
- 类型： 青春 励志 感人
- 我的评分：

★ ★ ★ ★ ★ ★ ★ ★ ★ ★

推荐电影

爱的捆绑
评分 7.3

假装有图片

少年手指虎
评分 7.2

时光网

欢迎您 身似 首页 电影列表 导演列表 演员列表 切换用户

垫底辣妹

总评分 8.1 打分

- 导演： 土井裕泰
- 主演： 有村架纯 伊藤淳史 吉田羊 田中哲司 野村周平
- 类型： 剧情 喜剧
- 地区： 日本
- 类型： 青春 励志 感人
- 我的评分：

★ ★ ★ ★ ★ ★ ★ ★ ★ ★

推荐电影

爱的捆绑
评分 7.3

少年手指虎
评分 7.2