**专业综合设计（大数据方向）课程报告**

2021-2022学年第1学期

|  |  |  |  |
| --- | --- | --- | --- |
| 项目评分表 | | | |
| 考察项目 | 评分项目 | 分值 | 得分 |
| 项目代码  （30分） | 满足项目功能需求，能以较快速度进行计算 | 15 |  |
|
| 代码结构合理，规范，易维护 | 15 |  |
|
| 项目答辩  （20分） | 解决方案的合理性 | 10 |  |
|
| 收获与体会 | 10 |  |
|
| 项目报告  （30分） | 问题分析与抽象，解决方案 | 5 |  |
| 设计系统结构，确定软硬件组成，确定系统边界 | 5 |  |
| 项目社会、健康、安全、法律、文化以及环境等可行性分析 | 4 |  |
| 技术规范标准适用分析，确定系统的主要技术指标 | 4 |  |
| 项目实施对法律、安全、健康、伦理与文化所产生的潜在影响分析及解决方案 | 4 |  |
| 团队成员职责分工与协调组织 | 4 |  |
| 软硬件选型的技术与经济性分析 | 4 |  |
| 项目总结  （10分） | 问题的抽象与分析 | 2 |  |
| 项目创新点说明与分析 | 2 |  |
| 技术指标说明，技术合理性分析 | 2 |  |
| 团队分工协作说明，成员完成任务情况说明 | 2 |  |
| 项目中工程管理方法与经济决策方法的使用说明 | 2 |  |
| 项目得分 |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 学生成绩表 | | | | |
| 学号 | 姓名 | 团队作用[[1]](#footnote-0) | 组内评分  （10分） | 合计成绩[[2]](#footnote-1) |
| 20194263 | 韦健俊 | 数据处理、集群 | 10 |  |
| 20194128 | 蒙家鑫 | 算法、集群 | 10 |  |
| 20194132 | 于杭渤 | 算法、数据处理 | 10 |  |
| 20194290 | 秦浩廷 | 前后端实现 | 10 |  |
| 评阅教师： | | | | |
| 评阅时间： | | | | |

|  |  |
| --- | --- |
| 专业综合设计任务书 | |
| 名称 | 电影推荐系统的设计与实现 |
| 类型 | □验证性 □设计性 综合性 |
| 内容 | （1）模仿MovieLens（https://movielens.org）构建一个电影评分网站；  （2）利用MovieLens Latest数据集（265MB，下载地址：http://files.grouplens.org/datasets/movielens/ml-latest.zip）作为网站的初始数据；  （3）采用推荐系统算法，计算用户偏好，根据用户的偏好向用户推荐电影；  （4）网站界面美观，电影推荐实时性较好，能以较少的内存运行。 |
| 要求 | （1）利用Docker或者云计算平台搭建分布式的数据处理集群，建议使用Spark作为大数据的处理引擎；  （2）推荐算法必须自己实现，并能在分布式数据处理平台上运行，**不允许**调用现有的软件库，**禁止使用**Spark中的推荐系统算法；  （3）数据存储使用NoSQL数据库，例如MongoDB；  （4）在处理大数据时，具有良好的性能，服务响应时间较短，并在报告中对算法的理论时间复杂度、空间复杂度和实际运行时间进行分析；  （5）推荐算法应具备处理MovieLens Latest数据集（265MB）的能力。 |
| 小组成员 | 20194263 计算机科学与技术3班 韦健俊  20194128 计算机科学与技术3班 蒙家鑫  20194132 计算机科学与技术3班 于杭渤  20194290 计算机科学与技术4班 秦浩廷 |
| 任务时间 | 2022年12月19日 |

电影推荐系统的设计与实现

1 项目背景介绍（于）

电影诞生于19世纪[[[3]](#endnote-0)]，起初是美国国家生活水平上升后，大众产生新需求的娱乐产物。随着近两个世纪的发展，电影已经成为了一种家喻户晓的视觉艺术作品。近年来，随着网络和数字媒体的高速发展，不断膨胀的用户数量和电影数目使得基于用户的“电影推荐系统”的实现非常重要。基于现有的用户观影数据，发掘用户的观影特征和各用户群体的相似度，从而为每位用户推荐可能符合其口味的新电影，对于选择困难症和没有明确需求的群体来说是一项尤为关键且实用的技术。

“电影推荐系统”的核心是推荐算法的实现，作为一项已经投入广泛应用的技术，国内外对于推荐算法的实现已经有了比较成熟的方案。推荐算法经历了从传统的协同过滤算法、因子分解机到深度学习方兴未艾的发展过程。尤其是近年来，随着一大批深度推荐模型的提出，深度学习模型成为了推荐系统的主流，效果越来越好。同时，海量数据的处理需求也催生了大数据平台工具的进化，大数据存储、传输、计算等组件发展日趋成熟[[[4]](#endnote-1)]。

总结国内外研究现状，电影推荐算法主要的实现方法基本分为两种：一种是近几年新兴的基于深度学习模型的推荐算法。深度学习模型的表达能力强，与传统模型相比能够进行特征间的深度交叉，从而挖掘出特征间更多的潜藏信息，网络结构也可以随着业务不同而改变，十分灵活。如利用现有数据集训练一个基于文本的卷积神经网络，或者更复杂的DIN，LSTM等深度学习模型作为排序层算法，采用Embedding技术作为快速召回算法，并利用TensorFlow Server建立模型服务，采用HDFS，Spark，Kafka，Flink等大数据存储等实现电影个性化推荐[[[5]](#endnote-2)]。另一种是传统推荐算法和基于其的创新，通过使用基于ALS算法的协同过滤推荐[[[6]](#endnote-3)]、基于用户的推荐、基于物品的推荐等方法中的一种或几种，或者将几种传统方法创造性转化、创新性发展，实现在网站界面中混合推荐的目的[[[7]](#endnote-4)]。

本项目为模仿MovieLens构建的一个电影推荐网站，利用MovieLens Latest数据集作为网站的初始数据，利用Docker搭建分布式的数据处理集群，使用Spark作为大数据的处理引擎。同时采用xxxxxx，计算数据特性、用户偏好，根据数据特性、用户的偏好推荐热门及用户个性化电影。本项目完整实现了前端可视化页面、后台业务处理、算法的设计与实现、环境的安装与部署等多种操作方式。网站界面美观，电影推荐实时性较好，能以较少的内存运行。

2 需求分析（蒙）

2.1功能性需求

1. 基本功能

Xxxxxxxxxxxxx。。。。。。。。。。

1. 推荐功能

xxxxxxxxxxxxxxx

1. 存储功能

Xxxxxxxxxxxx

2.2 非功能性需求

（1）项目可行性分析

Xxxxxxxxxxx

1. 项目实施的影响分析

本项目的实施符合国家安全、法律的要求，对健康、伦理等方面无不良影响。同时提供了浏览电影、了解电影的渠道，为用户推荐感兴趣的电影、为电影推荐潜在用户等功能具有实用性，对促进文化传播、增长用户视野方面起到积极作用。

1. 技术规范标准适用、软硬件选型的技术与经济性分析

Xxxxxxxxxxxxxx

3 系统设计与实现（自己分工）

3.1 总体架构设计（于）

自己画图

3.2 基本功能模块的设计与实现

3.2.1 数据加载模块（蒙）

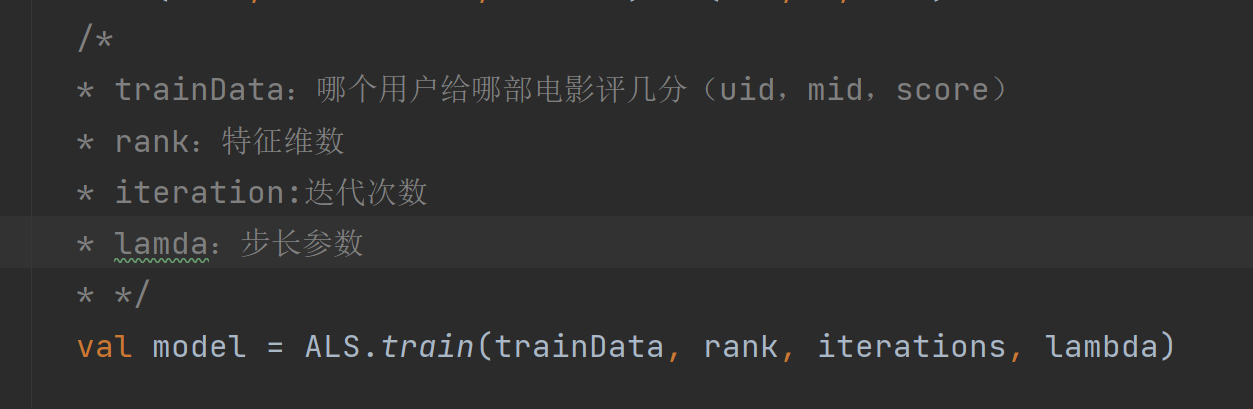
3.2.2 离线推荐模块（蒙）

离线推荐是根据历史数据如用户的评分来进行推荐，由于电影的评分会随着用户数和事件改变，所以离线推荐的算法模型参数并非固定，需要周期性的进行统计和训练。

这部分工作会由azkaban的调度来进行。

（1）首先读取mongodb中的数据，创建ALS训练模型。

读取ratings数据，得出uid，mid，score，形成三维特征值。然后放入ALS模型中进行训练。模型参数如下：



（2）

3.2.3 实时推荐（于）

3.2.4 综合业务服务（秦）

4 系统展示（韦）

展示界面、文字解释

5 总结

5.1 个人总结（每人一段）

5.2 项目总结（秦）

5.3 项目分工说明

|  |  |  |
| --- | --- | --- |
| 学号 | 姓名 | 工作 |
|  |  | Xxxxxxx |
|  |  | xxxxxxxxx |
|  |  | xxxxxxxxx |
|  |  | xxxxxxxx |

## 参考文献

1. 学生在团队中的作用由学生所在团队填写 [↑](#footnote-ref-0)
2. 合计成绩=项目得分+组内评分 [↑](#footnote-ref-1)
3. [↑](#endnote-ref-0)
4. [↑](#endnote-ref-1)
5. [↑](#endnote-ref-2)
6. [↑](#endnote-ref-3)
7. [↑](#endnote-ref-4)