|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | **卷 号** |  | | **卷内编号** |  | | **密 级** |  |   网易云音乐网站数据分析 |
| 分 类:  <模板>  使用者:  <项目管理部>  文档编号：  HD-PMC-305  四川华迪信息技术有限公司 | 例会纪要  项 目 承 担 部 门： 大数据1班05组  撰 写 人（签名）： 陈鹏宇  完 成 日 期： 2023-07-13  本文档 使 用部门： □主管领导 ☑项目组  □客户（市场） □维护人员 □用户  评审负责人（签名）： 陈鹏宇  会 议 日 期：2023-06-29至2023-07-13 共四次 |
| 华迪标志 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **会议记录** | | | | | |
| **日期**  **Date** | 2023年6月29日 | **时间**  **Time** | 15:00 | **地点**  **ADD.** | DS3304 |
| **主题：开题阶段安排** | | | | | |
| **主持人：陈鹏宇**  **出席者：徐小龙、杨奎、张一锋、付文韬、冯文康** | | | | | |
| **工作计划** | | **完成情况** | | **问题及解决方案** | |
| 1. 确定项目选题 2. 初步需求分析 3. 确定分工情况 | | 1. 确定项目选题：《网易云音乐网站数据分析》   完成度：100%   1. 初步需求分析，全组讨论选题需求和可实现的功能   完成度：90%   1. 确定分工情况   完成度：100% | | 问题1：如何项目管理、同步？  问题1解决方案：  统一使用github代码管理仓库。  问题2：如何分工，提高项目进度？  问题2解决方案：4个人负责数据爬取、清洗、处理的全流程，保证数据流通的顺畅，2个人负责看板设计、数据库设计等 | |

会议签到表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **姓 名** | **角 色** | **部 门** | **实到签名** | **备注** |
| 1 | 陈鹏宇 | 项目经理 | 大数据1班05组 | 陈鹏宇 |  |
| 2 | 徐小龙 | 编码人员 | 大数据1班05组 | 徐小龙 |  |
| 3 | 杨奎 | 编码人员 | 大数据1班05组 | 杨奎 |  |
| 4 | 张一锋 | 编码人员 | 大数据1班05组 | 张一锋 |  |
| 5 | 付文韬 | 编码人员 | 大数据1班05组 | 付文韬 |  |
| 6 | 冯文康 | 编码人员 | 大数据1班05组 | 冯文康 |  |
| 填表说明 | | | | | |
| 1. 角色可以是开发者、PPQA、评审员、记录员、评审组织者、管理人员、客户等，角色名应当反映在评审会议中担任的职责。 2. 本表由评审组织者在会议前准备好，并放置在会议室入口处，由参与人员签名。 | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **会议记录** | | | | | |
| **日期**  **Date** | 2023年7月3日 | **时间**  **Time** | 15:00 | **地点**  **ADD.** | DS3304 |
| **主题：第一阶段安排** | | | | | |
| **主持人：陈鹏宇**  **出席者：徐小龙、杨奎、张一锋、付文韬、冯文康** | | | | | |
| **工作计划** | | **完成情况** | | **问题及解决方案** | |
| 1. 爬取数据确定 2. 数据库设计方案 3. 第一阶段文档撰写 | | 1. 爬取数据确定   完成度：100%   1. 数据库设计方案   完成度：90%   1. 第一阶段文档撰写   完成度：100% | | 问题1： user-agent池中有少部分存在错误，导致网站返回内容形式与接收形式不兼容  问题1解决方案：禁用错误user-agent即可  问题2：程序间歇性被服务器识别为爬虫，导致接收到406状态码，造成大量（约1/5）数据丢失  问题2解决方案：为每次get/post请求添加retry功能，遭遇异常则进行循环等待请求，最大循环数设置为5次，每次等待时间设置为5s，有效降低了数据丢失率 | |

会议签到表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **姓 名** | **角 色** | **部 门** | **实到签名** | **备注** |
| 1 | 陈鹏宇 | 项目经理 | 大数据1班05组 | 陈鹏宇 |  |
| 2 | 徐小龙 | 编码人员 | 大数据1班05组 | 徐小龙 |  |
| 3 | 杨奎 | 编码人员 | 大数据1班05组 | 杨奎 |  |
| 4 | 张一锋 | 编码人员 | 大数据1班05组 | 张一锋 |  |
| 5 | 付文韬 | 编码人员 | 大数据1班05组 | 付文韬 |  |
| 6 | 冯文康 | 编码人员 | 大数据1班05组 | 冯文康 |  |
| 填表说明 | | | | | |
| 1. 角色可以是开发者、PPQA、评审员、记录员、评审组织者、管理人员、客户等，角色名应当反映在评审会议中担任的职责。 2. 本表由评审组织者在会议前准备好，并放置在会议室入口处，由参与人员签名。 | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **会议记录** | | | | | |
| **日期**  **Date** | 2023年7月06日 | **时间**  **Time** | 15:00 | **地点**  **ADD.** | DS3304 |
| **主题：第二阶段安排** | | | | | |
| **主持人：陈鹏宇**  **出席者：徐小龙、杨奎、张一锋、付文韬、冯文康** | | | | | |
| **工作计划** | | **完成情况** | | **问题及解决方案** | |
| 1. 清洗、处理数据方案 2. 页面设计方案 3. 第二阶段文档撰写 | | 1. 清洗、处理数据方案   完成度：100%   1. 页面设计方案   完成度：100%   1. 第二阶段文档撰写   完成度：100% | | 问题1：多进程爬虫开发中，起初循环遍历，为每一任务分配进程，但这会导致进程数量难以限制，造成程序/整机崩溃  问题1解决方案：改用Pool进程池，在创建时设定最大进程数即可  问题2：利用snownlp进行歌词和评论的emo指数分析时，发现snownlp的打分效果不佳  问题2解决方案：手动标注了一份约含2500条网易云音乐评论的小型语料库，重新训练了snownlp模型 | |

会议签到表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **姓 名** | **角 色** | **部 门** | **实到签名** | **备注** |
| 1 | 陈鹏宇 | 项目经理 | 大数据1班05组 | 陈鹏宇 |  |
| 2 | 徐小龙 | 编码人员 | 大数据1班05组 | 徐小龙 |  |
| 3 | 杨奎 | 编码人员 | 大数据1班05组 | 杨奎 |  |
| 4 | 张一锋 | 编码人员 | 大数据1班05组 | 张一锋 |  |
| 5 | 付文韬 | 编码人员 | 大数据1班05组 | 付文韬 |  |
| 6 | 冯文康 | 编码人员 | 大数据1班05组 | 冯文康 |  |
| 填表说明 | | | | | |
| 1. 角色可以是开发者、PPQA、评审员、记录员、评审组织者、管理人员、客户等，角色名应当反映在评审会议中担任的职责。 2. 本表由评审组织者在会议前准备好，并放置在会议室入口处，由参与人员签名。 | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **会议记录** | | | | | |
| **日期**  **Date** | 2023年7月13日 | **时间**  **Time** | 15:00 | **地点**  **ADD.** | DS3304 |
| **主题：第三阶段安排** | | | | | |
| **主持人：陈鹏宇**  **出席者：徐小龙、杨奎、张一锋、付文韬、冯文康** | | | | | |
| **工作计划** | | **完成情况** | | **问题及解决方案** | |
| 1. 页面逐步实现 2. 所有文档完善 3. 答辩准备 | | 1. 页面逐步实现   完成度：100%   1. 所有文档完善   完成度：100%   1. 答辩准备   完成度：100% | | 问题1： 花生壳https内网穿透后，生成的公共链接无法访问  问题1解决方案： 换成http服务  问题2：开启爬虫时，有时会出现大批量的爬取失败[retry overflow]  问题2解决方案：1.关闭代理 2.切换更优的网络 3.降低爬虫程序的进程数[在爬虫入口程序开头处] | |

会议签到表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **姓 名** | **角 色** | **部 门** | **实到签名** | **备注** |
| 1 | 陈鹏宇 | 项目经理 | 大数据1班05组 | 陈鹏宇 |  |
| 2 | 徐小龙 | 编码人员 | 大数据1班05组 | 徐小龙 |  |
| 3 | 杨奎 | 编码人员 | 大数据1班05组 | 杨奎 |  |
| 4 | 张一锋 | 编码人员 | 大数据1班05组 | 张一锋 |  |
| 5 | 付文韬 | 编码人员 | 大数据1班05组 | 付文韬 |  |
| 6 | 冯文康 | 编码人员 | 大数据1班05组 | 冯文康 |  |
| 填表说明 | | | | | |
| 1. 角色可以是开发者、PPQA、评审员、记录员、评审组织者、管理人员、客户等，角色名应当反映在评审会议中担任的职责。 2. 本表由评审组织者在会议前准备好，并放置在会议室入口处，由参与人员签名。 | | | | | |