**课程设计A**

**一、课程设计目的**

在Ubuntu 16.04(版本不限),Hadoop ,spark2.X或3.X环境下，巩固和加深对课程知识的理解，并能够综合利用所学知识，实现对数据的处理，尤其是解决在大规模分布式环境下的数据处理问题。

**二、数据来源**

使用老师给定的数据完成课程设计。

给定的数据集noaa-global-historical-climatology-network-daily介绍：

Noaa-global-historical-climatology-network-daily是全球历史气候网络-每日数据集，给定的数据是这个数据集的子集，涵盖了从1962年到2016年的气象数据。本次课程设计要求每位同选择一个年份（**同班同学之间不能重复**）的数据开展数据处理。

**三、课程设计**

以下统计分析、机器学习任务均应在Spark框架中完成，不利用Spark框架完成的项目作品将视为不符合题目要求(**可视化除外**)。

**1、数据预处理**

1) 分析数据集中每个字段含义，并给出中文描述

2) 分析数据集中不同文件之间的关联，给出描述（图形和文字）

3) 处理标题行（如果没有则需要添加标题行，不可以直接对原始数据进行编辑，只能在Spark中操作）

4) 检查是否有空值的项，并清理

5) 检查数据中是否有异常值（负值，非数值类型等）

6) 根据自己需要添加某些字段（比如时间可以拆分成年月日，方便后续统计）

7) 其它自定义可预处理项

**2、数据统计分析**

1) 显示数据集行数

2) 每个重要统计字段的不重复数据有哪些

3) 每个字段的最大、最小、平均值等信息（可统计字段）

4) 查询全年最大温度、最小温度、最大降雨、最小降雨的日期

5) 查询所用数据集是否有属于重庆的站点的数据，如果有则列出详细信息

6) 统计每月降雨量、最大的降雨量月份

7) 统计12月份的平均温度，最低温度，最高温度

**3、利用Spark机器学习库深入分析**

尽可能对数据进行仔细分析，利用所学的Spark机器学习库进行深入分析，利用机器学习方法发现规律。

利用机器学习算法对温度和降雨进行预测，可使用临近的年份为验证数据，预测方法可自行选择（但需要是Spark中的）,因天气本身也很难预测，因此重点不在于考查大家的预测准确率，而在于对工具的使用，以及对使用的方法的解释。

**4、数据可视化**

根据各自完成情况，将第2步“数据统计分析”中的2）、4）、6）以及第3步“利用Spark机器学习库深入分析”中得到的结果（可参考天预报的界面），利用所学的可视化方法进行可视化。

**四、提交课程设计包括的内容**

1、课程设计任务与目的；

2、课程设计完成过程（含关键程序源代码，及系统截图）；

3、心得体会；

4、参考文献；

5、附完整代码（放压缩包内）;

6、演示视频

**评分参考标准**

课程设计成绩分为本表成绩评定分优秀、良好、中等、及格、不及格五个等级。具体评分参考标准如下：

|  |  |
| --- | --- |
| **成绩等级** | **评估标准** |
| 优 | 对数据的预处理全面充分，能全部清理除数据中的异常数据。 使用Spark SQL或DataFrame 算子全面统计数据的内在特性，**能对输入的数据有较深的洞察并能在程序和文档中体现，能熟练使用**Spark机器学习库完成课程设计，数据**可视化结果具有多样性（截图与代码证明）**，**能充分展示数据分布的内在特性（并有结果解释）**。  提交的代码与截图结果**相符合，完全满足**课程设计题目要求。文档撰写格式规范，对课程设计过程描述详实，**每步骤均有对应详细描述**，课程总结全面且能全面真实反应课程设计过程中的问题及解决过程，**并能充分体现出个人对问题及解决的思考过程**。实现课程设计过程中能积极主动查阅相应文档**（尤其是官方文档）。** |
| 良 | 能使用RDD、DataFrame对数据进行预处理，能使用Spark SQL或DataFram完成统计数据分析，能使用Spark机器学习库完成机器学习分析任务，能够实现数据可视化。  提交的代码与截图结果**相符合**，**满足**课程设计题目要求。文档撰写格式规范，对课程设计过程描述**较详实**，课程总结**较充分**反应课程设计过程中的问题及解决过程。实现课程设计过程中能主动查阅相应文档，能够**较充分**理解算法基本原理，并完成课程设计。能够完成数据可视化。 |
| 中 | **基本能使用**RDD、DataFrame对数据进行预处理，存在部分异常未能清理的情况（或无对应清理代码），**基本能使用**Spark SQL或DataFram完成统计数据分析，统计分析覆盖面不全，基本能使用Spark机器学习库完成机器学习分析任务，但并不能充分满足课程设计要求，文档内容编写**基本符合**课程设计要求，能够进行数据可视化，但可视化不太全面。 |
| 及格 | **存在**数据预处理、Spark SQL或DataFram统计分析，Spark机器学习库、数据可视化四**者中一项未完成的情况**，但有提交文档且文档撰写**基本符合**课程设计要求，提交的代码可执行。 |
| 不及格 | **未能完成课程设计要求的内容（**存在数据预处理、Spark SQL或DataFram统计分析，Spark机器学习库、数据可视化四**者中2项及以上未完成的情况），文档撰写不符合要求**。  或者有下列情况之一者，课程设计成绩不及格：  **1）、未按要求完成课程设计；**  **2）、抄袭、剽窃他人结果。**  **3）、未提交相应代码** |