2024

金砖国家职业技能大赛（金砖国家未来技能和技术挑战赛）

机器学习与大数据

BRICS-FS-02-RU

样题（省级/区域选拔赛）

2024金砖国家职业技能大赛（金砖国家未来技能和技术挑战赛）

2024 年 02 月

目录

1 参赛形式

本次赛项为个人赛。

2 竞赛内容

本次竞赛由 3 个模块组成，要求参赛人员按顺序完成所有竞赛内容。竞赛时会向参赛人员提供统一的赛题文件、竞赛数据集、基础操作说明文件，以及为保障每个任务模块的独立性与公平性所需的技术基础条件。

竞赛内容包含基于机器学习与大数据的以下任务模块：

模块 A：大数据

模块 B：数据分析

模块 C：机器学习

如果参赛选手不遵守职业健康安全环境要求，或使自己和其他选手面临危险，他们可能会被取消比赛自核。

参赛选手完成模块后，将对结果进行评分。

3 项目模块和时间要求

3.1 项目模块和时间要求

机器学习与大数据赛项共 3 个模块，要求参赛者总共用时 360min。具体项目模块名称和时间要求参照表 1.

表 1 项目模块和时间要求清单

序号	模块名称	竞赛内容完成时间
1	模块 A：大数据	120min
2	模块 B：数据分析	120min
3	模块 C：机器学习	120min

3.2 任务内容

模块 A：大数据（120min）

模块描述：

Hadoop 得以在大数据处理应用中广泛应用得益于其自身在数据提取、变形和加载(ETL)方面上的天然优势。Hadoop 的分布式架构，将大数据处理引擎尽可能的靠近存储，对例如像 ETL 这样的批处理操作相对合适，因为类似这样操作的批处理结果可以直接走向存储。Hadoop 的 MapReduce 功能实现了将单个任务打碎，并将碎片任务(Map)发送到多个节点上，之后再以单个数据集的形式加载(Reduce)到数据仓库里 。

Sqoop 是一个分布式的数据迁移工具，可以将一个关系型数据库（例如 ：MySQL ,Oracle ,Postgres 等）中的数据导进到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导进到关系型数据库中。

hive 是基于 Hadoop 工具，用来进行数据提取、转化、加载，这是一种可以存储、查询和分析存储在 Hadoop 中的大规模数据的机制。hive 数据仓库工具能将结构化的数据文件映射为一张数据库表，并提供 SQL 查询功能，能将 SQL 语句转变成 MapReduce 任务来执行。

Spark 是一种与 Hadoop 相似的开源集群计算环境，但是两者之间还存在一些不同之处，这些有用的不同之处使 Spark 在某些工作负载方面表现得更加优越，换句话说，Spark 启用了内存分布数据集，除了能够提供交互式查询外，它还可以优化迭代工作负载。

任务一：Hadoop 基础操作

1、修改配置文件

在每个节点上操作，设置 hadoop.log.maxbackupindex 参数为 30，并保存修改完成后的配置文件。

2、启动 Hadoop 集群

检测是否成功启动 Hadoop 集群，包含 master 节点上的 NameNode、SecondaryNameNode 、ResourceManager，以及 slave1 、slave2 节点上的 NodeManager、DataNode，如果未启动，请启动 Hadoop 集群。

3、HDFS 创建目录

在 master 节点上操作，在 HDFS 上创建一个名为/BikeShare 的目录。

4、HDFS 上传文件

在 master 节点上操作,将~/BikeShare 文件夹里所有文件的数据（不包括第一行）写入到一个名为 bike_share.csv 的文件，并将 bike_share.csv 文件上传到 HDFS 的/BikeShare 目录下。

5、MapReduce 分区

在 master 节点上操作，打开 vscode-server 界面，切换根路径到/home/ec2-user/bike/（**必须切换根路径到指定文件夹，否则需选手自行配置 mapreduce 运行环境**），补全并运行 src/文件夹下的代码。读取 hdfs 上/BikeShare 目录下的数据，使用 mapreduce 程序统计各个自行车类型的骑行人数，同时每个自行车类型对应一个输出文件，输出路径为 hdfs 上的/BikeShareOutput 目录下。

任务二：Sqoop 数据迁移和 Hive 数据仓库

1、在 master 节点上操作，使用 Sqoop 将 slave1 节点上 Mysql 里 brics 数据库中 songs 表中数据全量导入到 Hive 数据仓库中的 default 数据库表 songs 中，字段类型为 string。

2、在 master 节点上操作，使用 Sqoop 将 slave1 节点上 Mysql 里 brics 数据库中 user_actions 数据表的全量数据导入到 hive 数据仓库中的 default 数据库表 user_actions 中，并实现 user_actions 表动态分区，Ds 为动态分区字段。

3、在 Hive 中的 default 数据库里创建具有以下特征的 bike_share 表，并将 hdfs 上/BikeShare 目录下的数据加载到 bike_share 表中。设置行分隔符为 \n，列分割符为，

4、处理 Hive 中的 bike_share 表，删除 started_at 字段中的"符号，并且

将 start_lat，start_lng，end_lat，end_lng 字段类型转换成 FLOAT 类型，最后将处理的结果表另存到 Hive 中 default.bike_share_clean 中。

5、探查 Hive 中的 bike_share_clean 表中数据，查询 member_casual 字段中的会员和非会员数据，分别计算两类人群平均每次骑行时长为多少分钟（结果只保存数值，且四舍五入取整），将结果按照要求格式保存到~/results/A_2_5.txt。

任务三：大数据实时采集

1、使用 ec2-user 用户完成 Zookeeper 和 Kafka 的配置安装，Zookeeper 和 Kafka 安装包地址/home/ec2-user/package。

2、在 master 节点上操作，实现 Flume 对 /home/ec2-user/flume/test.log 文件进行监控，将监控到的数据导入到 Kafka 的 test_topic 主题中。后台启动 flume-ng 服务，保证服务正常运行。

任务四：Spark 对音乐数据进行分析

1、编写代码，根据 Hive 中 user_actions 表统计每首歌的收藏人数，求收藏人数最多的歌曲 ID，并保存结果到~/results/A_4_1.txt 文件中。

2、读取 Hive 中 default 库里的 user_actions 表和 songs 表中数据，统计不同艺人每天的播放量。plays 字段作为艺人在当天的播放量，并将统计到的结果保存到 slave1 节点的 Mysql 中 brics 库中的 song_plays 表里。

模块 B：数据分析（120min）

模块描述：

通过对某平台用户浏览视频行为数据的研究，从用户、制作人、作品等维度

分析了不同指标，使用数据分析方法对浏览行为数据进行统计，得到不同主体的特征，进而了解不同主体的需求，并对不同主体的特征进行数据分析与挖掘，便于平台更好的开展业务，优化服务。

任务 1：数据读取与处理

任务 1.1 数据读取

运用 Pandas 读取`用户信息表`，`制作人信息表`和`作品信息表`三个表的数据,分别保存为变量`df_user`（用户信息表），`df_author`（制作人信息表）和`df_item `（作品信息表），并运行已给出的答案保存代码保存答案。

- `用户信息表`路径：`data/user.csv`

- `制作人信息表`路径：`data/author.xlsx`

- `作品信息表`路径：`data/item.txt`

例如：变量名 = pd.read_csv('路径')

任务 1.2 user 表处理

任务 1.2.1 无效列处理

对`df_user`变量进行处理，并运行已给出的答案保存代码保存答案，具体要求如下：

- 删除列字段`Unnamed: 0`

任务 1.2.2 字段类型修改

对`df_user`变量进行处理，并运行已给出的答案保存代码保存答案，具体要求如下：

- 修改列字段`uid`和`user_city`的数据类型为`object`。

任务 1.3 author 表处理

任务 1.3.1 重复值处理

对`df_author`变量进行处理，并运行已给出的答案保存代码保存答案，具体要求如下：

- 删除多余行重复值

任务 1.3.2 重命名列名）

对`df_author`变量进行处理，并运行已给出的答案保存代码保存答案，具体要求如下：

- 将字段`id`重命名为`author_id`

任务 1.4 item 表处理

任务 1.4.1 空值处理

对`df_item`变量进行处理，并运行已给出的答案保存代码保存答案，具体要求如下：

- 删除存在空值的行数据

任务 1.4.2 重置索引

对`df_item`变量进行处理，并运行已给出的答案保存代码保存答案，具体要求如下：

- 重置索引，并删除之前的索引

任务 1.4.3 时间字段处理

对`df_item`变量进行处理，并运行已给出的答案保存代码保存答案，具体要求如下：

- 提取`real_time`字段的日期信息，保存到新增列`date`。例如：2019-10-28

- 提取`real_time`字段的小时段信息，保存到新增列`hour`。例如：21；取值范围：0~23

任务 1.5 数据合并

对`df_user`、`df_author`、`df_item`三个变量进行处理，并运行已给出的答案保存代码保存答案，具体要求如下

- 将处理后的`df_user`、`df_author`、`df_item`3 个变量合并为一张表，并保存到变量`df`中。

- 3 张表以`item_id`字段进行内联接。

任务 2：数据分析

任务 2.1 用户数据分析

任务 2.1.1 总用户数量

分析`df_user`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 计算平台总用户数量，将答案保存在变量`task2_1_1`中。

任务 2.1.2 哪个城市观影用户最多？

分析`df_user`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 找出全国观影用户数量最多的城市`user_city`，将答案保存在变量`task2_1_2`中。

任务 2.1.3 不同渠道用户比例最高多少

分析`df_user`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 计算在不同渠道中，用户数量的最高占比（保留四位小数），将答案保存在变量`task2_1_3`中。

任务 2.1.4 有多少作品的看完率达到 100%?

分析`df_user`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 计算作品完播率达到 100%的作品数量。将答案保存在变量`task2_1_4`中。

作品完播率：指视频的播放完成率,即能够完整看完的人数比重

任务 2.1.5 被用户点赞过的作品数量有多少？

分析`df_user`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 计算被用户点赞过的作品数量。将答案保存在变量`task2_1_5`中。

任务 2.2 作者数据分析

任务 2.2.1 创作者数量

分析`df_author`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 计算平台制作人总数量，将答案保存在变量`task2_2_1`中。

任务 2.2.2 一部作品最多有多少人参与制作？

分析`df_author`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 在同一部作品中最多有几人参与制作？将答案保存在变量`task2_2_2`中。

任务 2.2.3 拥有作品最多的制作人是？

分析`df_author`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 在所有制作人中，拥有作品数量最多的制作人`author_id`是谁？将答案保存在变量`task2_2_3`中。

任务 2.2.4 拥有多部作品的制作人比例

分析`df_author`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 计算拥有多部作品的制作人数量占比（百分比形式，保留两位小数），将答案保存在变量`task2_2_4`中。

任务 2.3 作品数据分析

任务 2.3.1 总作品数量

分析`df_item`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 计算平台中的作品总数量，将答案保存在变量`task2_3_1`中。

任务 2.3.2 哪个城市是作品创作最热门城市

分析`df_item`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 找出作品创作最多的城市`item_city`，将答案保存在变量`task2_3_2`中。

任务 2.3.3 10 月份发布的作品中，内容涉及城市编号`4`，引用音乐 id 为`220`，总共有多少部作品？

分析`df_item`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 10 月份发布的作品
- 作品城市 id 为`4`的作品
- 音乐 id 为`220`的作品
- 计算满足以上三个条件的作品数量，将答案保存在变量`task2_3_3`中。

任务 2.3.4 集中在哪个时段发布作品

分析`df_item`变量，并运行已给出的答案保存代码保存答案，具体要求如下：

- 按以下时段区间进行分类

  0 <= 凌晨 < 6

  6 <= 上午 < 13

  13 <= 下午 < 19

  19 <= 晚上 < 24

- 计算哪个时段区间作品发布量最多？将答案保存在变量`task2_3_4`中。

任务 3：数据可视化

任务3.1 作品发布时间和发布作品数量图

统计作品发布数量随时间的变化趋势，请使用折线图进行展示，具体要求如下：

- x轴为作品发布时间，y轴为作品数量
- 折线图设置为点线类型
- x轴显示所有时间标签，并90°旋转

任务3.2 作品时长分布直方图

依据不同的作品内容时长统计作品数量，使用直方图进行展示，具体要求如下：

- 分析作品时长60秒以内（包括60）的数据
- x轴为作品内容时长，y轴为作品数量
- 同时显示直方图和核密度图

任务3.3 作品信息3D散点图

制作3D散点图，展示作品城市、作品音乐、作品时长之间的关系，具体要求如下：

- x轴为作品城市，y轴为作品音乐，z轴作品时长
- 不同颜色显示是否点赞，红色为点赞，反之为蓝色

模块 C：机器学习（120min）

模块描述：

该数据集描述的是葡萄牙银行的一个营销活动（定期存款），推广此次活动的主要手段是电话销售。通常为了让客户参加此次定期存款活动，需要不止一次联系客户。通过机器学习的方法，最终目是预测客户最后是否订阅了定期存款项目。

任务 1：数据探索

1.1 读取数据

读取`train`、`test`两个表的数据，将读取`train`表的所有数据保存到变量`data_train`中，读取`test`表的所有数据保存到变量`data_test`中，并运行给出的答案保存代码保存答案

例如：变量名 = pd.read_csv('表名.csv')

1.2 分析变量`data_train`，探查数据特征数量，求数据集一共有多少个特征，结果保存到变量`task1_2`中，并运行给出的答案保存代码保存答案

1.3 分析变量`data_train`,探查是否有缺失值的列,求有多少个列有缺失，结果保存到变量`task1_3`中，并运行给出的答案保存代码保存答案

1.4 分析变量`data_train`，探查数据标签分布，求标签为 0、1 各有多少个样本，结果保存到`task1_4`中，并运行给出的答案保存代码保存答案=

保存格式：(0 标签样本数，1 标签样本数)

任务 2 数据分析

2.1 分析变量`data_train`,假设此次活动最后一次联系才能确认客户的订阅意向，求哪一月的营销活动效果最好？结果保存到变量`task2_1`中，并运行给出的答案保存代码保存答案每月营销成功占比 = 每月订阅的客户人数/每月参与营销活动的客户人数

营销成功占比越高，营销活动效果越好

2.2 分析变量`data_train`，上一次未联系的客户中，成功参与本次活动订阅的用户占比为多少？（保留两位小数）结果保存到变量`task2_2`,并运行给出的答案保存代码保存答案

2.3 分析变量`data_train`中的`age`、`marital`字段，按照以下方式对年

龄进行分段，计算各年龄段的 married 人数，最多的是多少人？结果保存到变量`task2_3`中，并运行给出的答案保存代码保存答案

分段：

[0,10)：0<=age<10

[10,20)：10<=age<20

[20,30)：20<=age<30

[30,40)：30<=age<40

[40,50)：40<=age<50

[50,60)：50<=age<60

[60,70)：60<=age<70

[70,80)：70<=age<80

[80,90)：80<=age<90

[90,100)：90<=age<100

2.4 分析变量`data_train`，百分之八十的客户愿意与工作人员的通话时间只保持在几分钟内？结果保存到变量`task2_4`中，并运行给出的答案保存代码保存答案（2 分）

分钟数保留整数

任务 3.订阅意向预测

3.1 分析变量`data_train`、`data_test`，预测出`data_test`中字段 y 的结果，预测结果保存到变量`y_pred`，并运行给出的答案保存代码保存答案。

评分标准

目标

预测客户最后是否订阅了定期存款

评价指标

F1 分数：`sklearn.metrics.f1_score`

> F1 分数：是统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的精确率和召回率。F1 分数可以看作是模型精确率和召回率的一种调和平均，它的最大值是 1，最小值是 0。

最终得分

$$250 \times F1^3 - 36$$

注意

您应该提交一个包含 9043 行 2 列的 csv 文件。

该文件应该正好有 2 列： index（以 0 开始升序排序） y（包含您的二进制预测：1 订阅，0 表示未订阅）

4 项目模块评分标准

项目模块评分标准参照表 2。

表 2 评分标准

模块	任务	配分
A	大数据	30
B	数据分析	30
C	机器学习	40
合计		100

注：样题最终解释权归组委会所有。

2024全国职业技能大赛（全国未来技能和技术挑战赛）