深圳市第十二届职工技术创新运动会暨2022年深圳技能大赛

—大数据技术应用职业技能竞赛决赛样题

【主观题开始】

1、在Linux下安装Java并搭建完全分布式Hadoop集群。在Linux终端执行命令“initnetwork”，或双击桌面上名称为“初始化网络”的图标，初始化实训平台网络，启动hd-linux环境。

（1）在master节点下执行命令安装JDK1.8。对安装进度100%的结果进行截图。（1分）

（2）编辑“/etc/profile”文件，添加Java的环境变量，并且使配置生效。对添加至“/etc/profile”的内容进行截图。（1分）

（3）将JDK安装包发送至slave1与slave2节点，在slave1与slave2节点下安装JDK并且配置Java的环境变量并使配置生效。对添加至“/etc/profile”的内容进行截图。（1分）

（4）在master、slave1和slave2三个节点分别执行“java -version”，验证JDK安装成功并对输出结果进行截图。（2分）

（5）解压Hadoop安装包至“/usr/local”目录。解压后使用命令“ll /usr/local/”并对命令运行结果进行截图。（1分）

（6）修改Hadoop配置文件“core-site.xml”，设置HDFS的默认地址与端口（fs.defaultFS），设置HDFS的临时路径（hadoop.tmp.dir），对添加至“core-site.xml”的内容进行截图。（2分）

（7）修改Hadoop运行环境配置文件“hadoop-env.sh”以及YARN框架运行环境配置文件“yarn-env.sh”，添加Java的环境变量，并设置使用HDFS以及YARN框架的用户为root。分别对添加至“hadoop-env.sh”和“yarn-env.sh”的内容进行截图。（2分）

（8）修改HDFS配置文件“hdfs-site.xml”，设置Namenode数据存储的路径（dfs.namenode.name.dir）以及Datanode数据存储的路径（dfs.datanode.data.dir），设置HDFS文件块的副本数（dfs.replication）为2。对添加至“hdfs-site.xml”的内容进行截图。（1分）

（9）修改MapReduce配置文件“mapred-site.xml”，设置执行MapReduce作业的框架（mapreduce.framework.name）为YARN，设置MR App Master的环境变量（yarn.app.mapreduce.am.env、mapreduce.map.env和mapreduce.reduce.env），以及设置MapReduce作业所需的类路径CLASSPATH（mapreduce.application.classpath）。对添加至“mapred-site.xml”的内容进行截图。（5分）

（10）获取已完整配置的“yarn-site.xml”，复制到Hadoop对应的配置文件目录并替换。替换完成后打开“yarn-site.xml”对文件内容进行截图。（1分）

（11）修改workers文件，设置为slave1与slave2。对workers的文件内容进行截图。（1分）

（12）发送master中的Hadoop至slave1与slave2节点。发送成功后分别在slave1与slave2节点上执行“ll/usr/local/”，并对命令的运行结果进行截图。（2分）。

（13）在三个节点的“/etc/profile”添加Hadoop路径（HADOOP\_HOME），并且使配置生效。对master、slave1、slave2三个节点添加至“/etc/profile”的内容进行截图。（1分）

（14）格式化NameNode并启动Hadoop集群（包括HDFS、YARN和historyserver）。对格式化NameNode的结果进行截图（即显示“successfully formatted”字样），并且对启动Hadoop集群的命令执行结果进行截图。（2分）

（15）在master、slave1和slave2三个节点分别执行“jps”，验证Hadoop集群启动成功并且对输出结果进行截图。（2分）

【数据获取】

（1）使用wget命令获取JDK安装包：

“wget -P /data http://house.tipdm.com/SZ-Competition/jdk-8u281-linux-x64.rpm”

1. 使用wget命令获取Hadoop安装包：

“wget -P /data http://house.tipdm.com/SZ-Competition/hadoop-3.1.4.tar.gz”

1. 使用wget命令获取yarn-site.xml：

“wget -P /data http://house.tipdm.com/SZ-Competition/yarn-site.xml”

（运行引号中的所有代码可将数据上传到实训平台）

【文件读取路径】

“/data/jdk-8u281-linux-x64.rpm”

“/data/hadoop-3.1.4.tar.gz”

“/data/yarn-site.xml”

【答案提交】

最终答案需将实现命令与运行截图整理成一份文档，并以“准考证号（20220821XXX）+问题1”命名进行提交

【主观题开始】

2、请依据题目要求运用Python语言采集网站中二手房的信息。

第一页网址：https://lz.esf.fang.com/house/i31/?rfss=1-a9b02b3f95ca4c2c75-74

第二页网址：https://lz.esf.fang.com/house/i32/?rfss=1-b8aa0929c900073556-36

某间二手房网址：https://lz.esf.fang.com/chushou/3\_451967928.htm?rfss=1-a9b02b3f95ca4c2c75-74

（1）解析房天下网站二手房的第一页与第二页网址，采集每间二手房网址，并将采集结果存于Excel文件，保存路径设定为“/data/result2\_1.xlsx”。（5分）



（2）对比不同网址信息的分布结构，采集二手房信息的总价、户型、朝向、建筑面积、楼层、单价与装修情况，并将数据整理为数据框形式。（13分）



（3）将数据以“['总价','户型','朝向','建筑面积','楼层','单价','装修']”中文名称设定列名，最终将数据保存为Excel文件，保存路径设定为“/data/result2\_2.xlsx”。（2分）

【答案提交】

提交代码文件（.ipynb格式文件和.html格式文件）和采集数据（result2\_1.xlsx和result2\_2.xlsx）。文件使用压缩包的形式上传，并以“准考证号（20220821XXX）+问题2”命名

【主观题开始】

3、基于采集的二手房信息数据，根据题目要求运用Spark框架与Scala语言编程（使用IntelliJ IDEA）实现二手房数据清洗与挖掘。（文件名及路径为“/data/SecondhandHouseV1.csv”）。

（1）上传“SecondhandHouseV1.csv”至HDFS的“/user/root”目录。上传完成后使用hdfs命令查看该目录并截图。（2分）

（2）结合Spark框架，在IDEA编写Scala代码，读取HDFS中的二手房屋数据集“SecondhandHouseV1.csv”为DataFrame，查看数据集的前6条数据并对结果截图。（2分）

（3）用null值替换数据中含有“暂无”两字的信息，再删除含有null的行数据。统计原数据集的行列数以及删除null值后数据集的行列数，对输出结果截图。（3分）

（4）统计行数据重复的数量，并将结果进行展示；对重复的行数据进行删除，统计删除重复数据后数据集的行列数，对输出结果截图。（3分）

（5）删除“建筑年代”列的年份单位“年”，仅保留数值，并将其数据类型转为整数型。（2分）

（6）二手房的出售一般为旧楼房，建筑年代应小于2022年，保留“建筑年代”小于等于2021年的房屋数据。展示处理后数据集的行列数，对输出结果截图。（2分）

（7）使用当前年份（2022年）减去建筑年份获取房龄，并将结果存入“房龄”列。（2分）

（8）删除“户型”、“建筑年代”和“住宅类别”列。（2分）

（9）完成上述步骤后，将处理后的二手房屋数据集DataFrame以csv格式导出至“/data/result3”目录。查看csv文件并截图。（2分）

【数据获取】

使用wget命令获取SecondhandHouseV1.csv：

“wget -P /data http://datasrc.tipdm.net:81/bigdata/SocialCompete/SecondhandHouseV1.csv”

（运行引号中的所有代码可将数据上传到实训平台）

【文件读取路径】

“/data/SecondhandHouseV1.csv”

【答案提交】

最终答案需将实现命令与运行截图整理成一份文档，并与采集数据（result3.csv）放于同一文件夹。文件使用压缩包的形式上传，并以“准考证号（20220821XXX）+问题3”命名进行提交。

【主观题开始】

4、基于二手房房价数据，根据题目要求运用数据挖掘与可视化知识对数据进行统计与基本图形绘制。（文件名及路径为“/data/SecondhandHouse\_view.csv”）

（1）绘制建筑面积与房价分布情况的散点图：设定x轴数据为建筑面积，y轴数据为总价；x轴与y轴标签（xlabel and ylabel）分别为“建筑面积（平米）”和“总价（万）”；图形标题（title）为“二手房建筑面积与房价的关系分析”。（4分）

（2）运用seaborn库绘制不同装修程度的二手房房价的分组箱线图：设定x轴数据为装修程度，y轴数据为总价；x轴与y轴标签（xlabel and ylabel）分别为“装修程度”和“二手房房价（万）”；图形标题（title）为“不同装修程度的二手房房价分组箱线图”。（5分）

（3）统计“楼层”、“电梯”和“学校”列不同情况的二手房数量，并进行打印展示。（3分）

（4）绘制二手房楼层、电梯和学校的不同情况占比的饼图：将画布分成1行3列的3个子图，子图1、2、3分别绘制楼层、电梯和学校不同情况占比的饼图；对每个子图设定标题，分别为“楼层”、“电梯”和“学校”；令每个饼图展示各自的百分比（autopct）和标签（labels），其中百分比保留小数点后1位。（8分）

【数据获取】

方法①使用wget代码获取数据：

“!wget -P /data http://datasrc.tipdm.net:81/python/testdata/ShenZhen-JS/SecondhandHouse\_view.csv”

（运行引号中的所有代码可将数据上传到实训平台）

方法②下载题目附件中的数据，上传到实训平台中

【文件读取路径】

“/data/SecondhandHouse\_view.csv”

【答案提交】

最终答案需提交代码文件（.ipynb格式文件和.html格式文件），文件使用压缩包的形式上传，并以“准考证号（20220821XXX）+问题4”命名

【主观题开始】

5、基于兰州二手房房价数据，根据题目要求运用机器学习知识实现数据建模与评估。

（文件名及路径为“/data/SecondhandHouse\_view.csv”）

（1）特征编码：将“朝向”、“楼层”、“装修”、“电梯”、“产权性质”、“建筑结构”、“建筑类别”、“区域”和“学校”9列数据的类型由字符型转化为数值型，如“电梯”列，原{'有','无'}转化为{1,0}。特征编码后的数据保存为CSV文件，保存路径设定为“/data/result5.csv”。（9分）

（2）数据集划分：提取“总价”列作为标签，其他列作为特征；按8:2的比例划分训练集与测试集，并展示划分后训练集与测试集的数据性质（shape）。（2分）

（3）模型训练与评估：使用sklearn估计器构建回归模型，放入训练集进行训练；放入测试集特征实现模型预测；结合预测结果与测试集标签，运用klearn.metrics模块的r2\_score函数评估模型优劣，并展示评估结果。（本小题主要考察模型构建与评估，因此R方值的大小影响题目得分，越接近1表明模型越好，该题得分越高，4分）

【数据获取】

方法①使用wget代码获取数据：

“!wget -P /data http://datasrc.tipdm.net:81/python/testdata/ShenZhen-JS/SecondhandHouse\_view.csv”

（运行引号中的所有代码可将数据上传到实训平台）

方法②下载题目附件中的数据，上传到实训平台中

【文件读取路径】

“/data/SecondhandHouse\_view.csv”

【答案提交】

最终答案需提交代码文件（.ipynb格式文件和.html格式文件）和结果数据（result5.csv），文件使用压缩包的形式上传，并以“准考证号（20220821XXX）+问题5”命名。