# 2025 年河北省职业院校大数据应用与服务（中职组）赛项样题

## 一、背景描述

近年来，随着旅游业的快速发展和社交媒体的普及，一些目的地因其独特的魅力或者事件而迅速走红，吸引了大量游客涌入，使得当地的酒店业面临着客房需求相应上升。在需求旺盛的时候，酒店为了保持利润往往会提高房价。大数据时代背景下，为避免游客到达旅游地之后因住宿产生高额费用且觉得酒店并不适合自己提供一种全新的思路，游客在出行前可以做好详细的攻略，提前预定最适合自己的酒店。性价比作为决定选择哪一家酒店的最重要的因素。通过对大量酒店信息数据进行分析研究，可以对酒店进行比较准确客观的评分和评价，或者进行相应的用户画像，游客可以选择适合自己的酒店进行相应的预定。

因数据驱动的大数据时代已经到来，没有大数据，我们无法为用户提供大部分服务，为完成互联网酒店的大数据分析工作，你所在的小组将应用大数据技术，通过 Python 语言 以数据采集为基础，将采集的数据进行相应处理，并且进行数据标注、数据分析与可视化、通过大数据业务分析方法实现相应数据分析。运行维护数据库系统保障存储数据的

安全性。通过运用相关大数据工具软件解决具体业务问题。你们作为该小组的技术人员， 请按照下面任务完成本次工作。

## 二、竞赛内容

竞赛分模块1、模块2和模块3三个部分，详见下表：

**表1竞赛内容说明**

| 模块 | 任务 | 分数 | 竞赛时间 |
|---|---|---|---|
| 模块 1：平台搭建与运维 | 任务 1：大数据平台搭建 | 10 | 240 分钟 |
| | 任务 2：数据库配置维护 | 20 | |
| 模块 2：数据采集与处理 | 任务 1：数据获取与清洗 | 10 | |
| | 任务 2：数据标注 | 10 | |
| | 任务 3：数据统计 | 15 | |
| 模块 3：业务数据分析与可视化 | 任务 1：数据分析与可视化 | 20 | |
| | 任务 2：业务分析 | 10 | |
| | 考察职业素养 | 5 | |

## 三、竞赛任务提交

参赛选手按照三个模块的任务要求完成对应的答题后提交验证。

## 四、竞赛注意事项

提交的答题记录内容中，不能填写与参赛选手相关的信息，如姓名和院校。如出现上述标记，成绩按照零分处理。

## 五、模块1：平台搭建与运维

### （一）任务1：大数据平台搭建

1.子任务1-1：Hadoop完全分布式安装配置

（1）在master将jdk-8u212-linux-x64.tar.gz、hadoop-3.1.3.tar.gz解压到/root/software目录下。

（2）在master上生成SSH密钥对，实现三台机器间的免密登录。并同步jdk，配置环境变量并生效。

（3）将hadoop分发至slave1、slave2中，其中三个节点节点均作为datanode，配置好相关环境，初始化Hadoop环境namenode。

（4）开启集群，查看各节点进程。

2.子任务1-2：MySQL安装配置

（1）解压MySQL 5.7.25到/root/software目录下，并安装MySQL组件。安装好MySQL后，使用mysql用户初始化和启动数据库。

（2）无密码登录MySQL，并修改root用户的密码为123456，验证登录。

（3）增加用户远程登录权限。

4.子任务2-2：Hive安装配置

（1）将Hive3.1.2安装包解压到/root/software目录下；并配置其环境变量，让其立即生效，查看Hive版本；

（2）修改相关配置，添加依赖包，将MySQL数据库作为Hive元数据库，初始化Hive元数据。

**（二）任务2：数据库配置维护**

1.子任务2-1：创建数据库及相关数据表

在 MySQL 数据库中创建“test”数据库，并在“test”数据库中分别创建“shopping”、“fooditems”共 2 个数据表。各个数据表的表字段格式如下；

**表2：fooditems表**

| 字段 | 字段中文名 | 类型 | 备注 |
|---|---|---|---|
| id | 行号 | INT | 自增，主键 |
| city | 城市 | VARCHAR(255) | |
| food_name | 美食名称 | VARCHAR(255) | |
| likelihood_of_liking | 喜爱度 | INT | |
| restaurant_list | 餐馆列表 | TEXT | |
| food_detail_link | 美食详情链接 | TEXT | |
| food_image_link | 美食图片链接 | TEXT | |
| food_description | 美食介绍 | TEXT | |

**表3：shopping表**

| 字段 | 字段中文名 | 类型 | 备注 |
|---|---|---|---|
| id | 行号 | INT | 自增，主键 |
| city | 城市 | VARCHAR(255) | |
| shop_name | 购物地名称 | VARCHAR(500) | |
| address | 地址 | VARCHAR(50) | |
| contact_phone | 联系电话 | VARCHAR(100) | |
| business_hours | 营业时间 | VARCHAR(100) | |
| ranking | 排名 | VARCHAR(100) | |

| | | | |
|---|---|---|---|
| overall_rating | 综合评分 | VARCHAR(50) | |
| reviews_count | 点评数 | VARCHAR(50) | |
| review_category | 评价类别 | VARCHAR(100) | |
| visitor_rating | 游客评分 | VARCHAR(100) | |
| visitor_review | 游客评价 | TEXT | |

（1）参考以上字段信息创建“test”数据库、“shopping”表、“fooditems”表。

2.子任务2-2：添加数据记录

（1）将本地/root/shopping/目录下的数据文件shopping.csv导入MySQL对应数据库shopping数据表；

（2）将本地/root/food/目录下的数据文件fooditems.csv导入MySQL对应数据库fooditems；

3.子任务2-3：维护数据表

结合已导入的两份sql数据，对其中的数据进行如下查询和操作。

（1）对test’数据库中的‘shopping’数据表进行查询，查询游客对于果戈里书店的环境的评分，字段名称命名为'环境评分'，将结果保存至view_table01中；

（2）对‘test’数据库中的‘fooditems’数据表进行查询，查询表中北京有多少种美食，字段命名为‘美食个数’，将结果保存至view_table02中；

（3）对‘test’数据库中的‘shopping’数据表进行查询，查询表中综合评分4.5以上，且排名在前一百名之内

的店铺有几个，字段命名为'个数'，将结果保存至view_table03中；

（4）对‘test’数据库中的‘fooditems’数据表进行查询，查询有美食‘麻豆腐’的城市有哪些，将结果保存至view_table04中；

## 六、模块二：数据获取与处理

### （一）任务1：数据获取与清洗

#### 1.子任务1-1：数据获取

（1）有一份酒店详情列表数据：酒店名称、酒店类型、位置信息、起价、评分、点评数，并且存入到 hotel.txt 文件中。使用 pandas 读取 hotel.txt 并将读取的txt文件前十行。

（2）有一份酒店综合评价数据：城市、酒店名称、酒店类型、地址、报价、酒店排名、综合评分、点评数、最热评价，并且存入到 hotels.txt 文件中。使用 pandas 读取 hotels.txt 并将读取的txt文件前十行。

#### 2.子任务1-2：数据清洗

现已从相关网站及平台获取到原始数据集，为保障用户隐私和行业敏感信息，已进行数据脱敏。数据脱敏是指对某些敏感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。在涉及客户安全数据或者一些商业性敏感数

据的情况、不违反系统规则条件下，对真实数据进行改造并提供测试使用，如身份证号、手机号等个人信息都需要进行数据脱敏。

相关数据文件中已经包含了数据采集阶段从某旅游平台网站上爬取的数据集，其中包含了来自哈尔滨的酒店信息，你的小组需要通过编写代码或脚本完成对相关数据文件中住宿场所销售管理数据的清洗和整理。

请使用 pandas 库加载并分析相关数据集，根据题目规定要求使用 pandas 库实现数据处理，具体要求如下：

（1）删除 hotel.txt 中酒店类型为空的数据并且存入hotel2_c1_N.csv,N 为删除的数据条数；

（2）去除 hotel.txt 中起价中除数字外的其他字符，生成新数据列‘最低价’，并且存入 hotel2_c2.csv；

（3）将 hotel.txt 中评分为空的数据设置为 0 并且存入 hotel2_c3.csv；

（4）将 hotel.txt 中评分为空的数据设置为总平均评分保留一位小数并且存入 hotel2_c4_N.csv，N为总平均评分保留一位小数；

（5）删除 hotels.txt 中最热评价为空的数据并且存入 hotel_comment.csv。

（二）任务2：数据标注

1.子任务2-1：分类标注

使用 SnowNLP 对酒店评论数据 hotel_comment.csv 进行标注， 获取情感倾向评分（sentiments）,具体的标注规则如下：

（1）对情感倾向分数大于等于 0.7 评论数据标注为正向；

（2）对情感倾向分数大于 0.4 小于 0.7评论数据为中性；

（3）对情感倾向分数小于等于 0.4 评论数据标注为负向。

根据采集到的评论信息，给出三类标注好的数据，存入 standard.csv。具体格式如下：

| 编号 | 酒店名称 | 评论信息 | 情感倾向 | 备注 |
|---|---|---|---|---|
| 1 | 亚季酒店(哈尔滨太平国际机场店) | XXXXXX | 正向 | |

（三）任务3：数据统计

1.子任务2-1：HDFS文件上传下载

本任务需要使用 Hadoop 、HDFS 命令，已安装 Hadoop 及需要配置前置环境，具体要求如下：

（1）在 HDFS 目录下新建目录/file2_ 1，查看目录；

（2）修改权限， 赋予目录/file2_ 1 最高 777 权限；

（3）下载 HDFS 新建目录/file2_ 1，到本机指定目录/root/下。

2.子任务2-2：处理异常数据

（1）处理异常数据

hotel.txt 文件存储了旅游平台网站上收集的用户哈尔滨酒店数据，数据中有以下字段：酒店名称、酒店类型、位置信息、起价、评分、点评数。

编写 Python 代码或程序，实现以下功能：将 hotel.txt 数据中位置信息字段中分隔符“·”替换为“,”。并以“,”为分隔符将位置信息字段分隔为商圈与景点字段，并将分割后的数据输入到/root/district.csv文件中，然后在控制台按顺序打印输出前 10 条数据。

3.子任务2-3：数据统计

（1）Python进行数据统计

district.csv文件存储了旅游平台网站上收集的用户哈尔滨酒店数据，数据中有以下字段：酒店名称、酒店类型、位置信息、商圈、景点、起价、评分、点评数。

编写 Python 代码或程序，实现以下功能：分析 district.csv 数据中商圈前三的酒店类型有哪些，并将输出的数据写入到/root/types.csv文件中，然后在控制台按顺序打印输出前 10 条数据。

（2）MapReduce程序进行数据统计

District_etl.csv文件存储了旅游平台网站上收集的用户哈尔滨酒店数据，数据中有以下字段：酒店名称、酒店类型、位置信息、商圈、景点、起价、评分、点评数。

编写 MapReduce 程序，实现以下功能：分析 district_etl.csv 数据中不同评分区间（4.0分以下、大于等于4.0分-小于4.5分、大于等于4.5分-小于等于5.0分）的酒店有多少，将结果写入 HDFS，在控制台读取 HDFS 文件。

## 七、模块3：业务数据分析与可视化

### （一）任务1：数据分析与可视化

1.子任务1-1：使用 Python 进行数据分析和可视化

数据分析

游客出行做攻略的首要目标便是选择适合自己的酒店，其中价格、位置、酒店类型便是游客做决定的重要因素。请编写程序或脚本根据模块二任务三中子任务二处理完成得到的数据文件 district.csv 统计以下的相关信息：

（1）分别统计各个商圈的的酒店总数，进行倒序排序展示前五名；

（2）统计各个商圈酒店的平均最低价，进行正序排序展示前五名；

（3）统计所有五星级酒店的平均评分。

数据可视化

在企业消费平台上， 各地区的酒店信息能够反映一个地 区商业活动的密集程度。例如酒店总量多的城市大都具有强 烈的吸纳外来人员的能力，订单数量能够反映该地区的有较多的商业往来。根据现有数据及给定参数完成酒店数据统计。

使用 Python 代码编写数据可视化的相关功能，所用数据为模块二任务三中子任务二处理完成得到的数据文件 district.csv 数据。

（4）用柱状图显示商圈的酒店总数排名前十的商圈；

（5）用折线图显示各类型酒店平均评分走势。

2.子任务1-2：Echarts可视化

ECharts.js是一款基于HTML5的图形库。图形的创建也比较简单，直接引用Javascript即可。

（1）结合数据统计中的MapReduce程序运行结果，引入ECharts文件，并使用饼图方式进行可视化展示；

（2）结合数据分析中的商圈酒店数量排名前五的运行结果，引入ECharts文件，并使用柱状图进行可视化展示；

（3）结合数据统计中的python程序运行结果，引入ECharts文件，并使用簇状柱形图方式进行可视化展示。

（二）任务2： 业务分析与方案设计

1.子任务2-1：业务分析

完成模块二任务二已标注数据 standard.csv 评论情感分析功能，酒店的正向、中性、负向评价数量，绘制柱状图，并对酒店的发展趋势作出简要分析。

2.子任务2-2：报表分析

根据模块二任务三中子任务二处理完成得到的数据文件 district.csv ，通过 Excel 生成报表信息方便游客在攻略服务中进行预定经济整洁，及时准确的把握不同类型酒店的评分变化，根据酒店评分中的4.8、4.9与5.0的舒适型酒店占比绘制饼状图。