# 2023 年四川省职业院校技能大赛（中职组）
# 大数据应用与服务赛项样题

## 背景描述

大数据时代背景下，人们生活习惯发生了很多改变。在传统运营模式中，缺乏数据积累，人们在做出一些决策行为过程中，更多是凭借个人经验和直觉，发展路径比较自我封闭。而大数据时代，为人们提供一种全新的思路，通过大量的数据分析得出的结果将更加现实和准确。平台可以根据用户的浏览，点击，评论等行为信息数据进行收集和整理。通过大量用户的行为可以对某一个产品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，将产品推荐给喜欢该产品的用户进行相应的消费。

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，并且进行数据分析与可视化、数据标注、通过大数据业务分析方法和方案架构实现相应应用功能。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。

# 任务 A：数据采集与处理

## 子任务一：数据采集

网站解析，利用 Chrome 查看网页源码，分析企业消费平台网站网页结构。

1.打开企业消费平台网站，在网页中右键点击检查，或者 F12 快捷键，查看元素页面；

2．检查网站：浏览网站源码查看所需内容。

从企业消费平台网站中爬取需要数据，按照要求使用 Python 语言编写爬虫代码，爬取指定数据项，并对结果数据集进行数据探索、以及必要的数据处理操作。请将符合题目要求的代码答案复制粘贴至对应报告中。

具体步骤如下：

（1）创建爬虫项目

（2）构建爬虫请求

（3）按要求定义相关字段

（4）获取有效数据

（5）将爬取到的数据保存到指定位置

至此已从企业消费平台网站中爬取了所需数据，下一步我们要将爬取结果进一步进行相关数据操作。具体要求如下：

爬取酒店列表数据，例如酒店名称、国家、省份、城市、商圈、星级、房间数、图片数、评分、评论数并且存入到 hotel.csv 文件中。

## 子任务二：数据处理

1.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。

2.相关数据文件中已经包含了数据采集阶段从企业消费平台网站上爬取的数据集，其中包含了来自不同城市的多家住宿场所的销售信息，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。

3.请分析相关数据集，根据题目规定要求实现数据处理，具体要求如下：

4.删除 hotel2.csv 中酒店名称为空的数据并且存入 hotel2_c1.csv；

5.删除 hotel2.csv 中删除数据源中缺失值大于 3 个字段的数据记录并且存入 hotel2_c2.csv；

6.将 hotel2.csv 中评分为空的数据设置为 0 并且存入 hotel2_c3.csv；

7.将 hotel2.csv 中评分为空的数据设置为平均评分并且存入 hotel2_c4.csv。

## 任务 B：数据分析与可视化

### 子任务一：数据分析

1.城市游客接纳能力是城市规划建设中的重要指标，其中城市的酒店房间数量是城市游客接纳能力的关键要素。请编写程序或脚本根据酒店管理网站中的数据 hotel_all.csv 统计以下的相关信息，具体要求如下：

2.分别统计北京、上海、广州、深圳的酒店总数；

3.统计北京、上海、广州、深圳所有酒店的平均评分排名；

4.统计上海酒店的平均房间数；

5.统计所有 5 星级酒店的平均评分。

### 子任务二：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，数据分析业务所用数据为 hotel_all.csv 数据，具体要求如下：

用柱状图显示北京、上海、广州、深圳酒店总数；

用折线图显示北京、上海、广州、深圳 4 星级酒店平均评分走势；

用饼图显示北京各星级酒店数占比。

## 任务 C：数据标注

### 子任务一：分类标注

对酒店评论数据 hotel_comment.csv 进行标注，具体的标注规则如下：

1.对具有想想反馈的评论数据标注为正向；

2.对不具备情绪反应的数据标注为中性数据，如毫无意义的灌水评论等；

3.对批判、讽刺等具有负向反馈的评论信息标注为负向。

4.根据采集到的评论信息，给出三类标注好的数据，每个类型 100 条，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 | 全季酒店 | XXXXXX | 中性 | |
| | | | | |

## 任务 D：业务分析和方案架构设计

### 子任务一：业务分析

完成 hotel_comment_all.csv 评论情感分析功能，以月度为单位统计每月某酒店的正向、负向评价数量，绘制折线图，并对酒店的发展趋势作出简要分析。

### 子任务二：报表分析

1.根据已标注数据 standard_c1.csv 文件中的结果，通过 excel 生成

报表信息方便产品方在后续作品中进行优化，及时准确的把握市场行情，具体要求如下：

2.某酒店的评论正向和负向的评论区趋势；

3.某酒店在互联网上的整体评价趋势；

4.某酒店正向评论前 5 个词及负向评论的前 5 个词。

## 任务 E：数据库维护

### 子任务一：创建相关表

1.根据采集数据字段在 MySQL 数据库中创建酒店表（hotel）。酒店表字段如下：

| 字段 | 类型 | 中文含义 | 备注 |
|---|---|---|---|
| id | int | 酒店编号 | |
| name | varchar | 酒店名称 | |
| city | varchar | 城市 | |
| star | int | 星级 | |
| room_num | int | 房间数 | |
| image_num | int | 图片数 | |
| score | double | 评分 | |
| comment_num | int | 评论数 | |

2.根据采集数据字段在 MySQL 数据库中创建评论表（comment）。评论表字段如下：

| 字段 | 类型 | 中文含义 | 备注 |
|---|---|---|---|
| id | int | 评论编号 | |
| name | varchar | 酒店名称 | |
| commentator | varchar | 评论人 | |

| | | | |
|---|---|---|---|
| score | double | 评分 | |
| comment_time | datetime | 评论时间 | |
| address | varchar | 评论位置 | |
| content | varchar | 评论内容 | |

## 子任务二：维护数据表

1.在 hotel 表中删除 id 为 25 的酒店数据;

2.在 comment 表中将 id 为 30 的评论数据地址改为北京;

3.统计北京、上海、广州、深圳的酒店总数。