浙江工业大学

人工智能原理及应用
实验报告

实验名称：《披荆斩棘的哥哥》数据爬取与分析

学　　号：202205240220

姓　　名：潘家航

专业班级：自动化 2304

学　　院：信息工程学院

指导教师：付明磊

目录

一、实验目的

通过 Python 编程实现对百度百科网页的自动化数据爬取，提取综艺节目《披荆斩棘的哥哥》参赛选手的基本信息（姓名、出生年份、身高、体重等），并利用 pandas 和 matplotlib 对数据进行统计与可视化分析，从而熟悉网络数据采集与可视化的基本流程。

二、实验设备

1.硬件设备： 游侠 G15。

2.软件环境： 主流浏览器（如 Chrome、Edge）、代码编辑器 pycharm、Python 编程环境。

3.开发平台： 百度 AI 开放平台（ai.baidu.com）的 EasyDL 产品。

三、实验内容

1.网络爬虫基本流程：

发送 HTTP 请求获取网页源代码，解析 HTML 文档结构提取有效信息（如姓名、出生年份、身高、体重等），将结果保存为结构化数据格式（如 CSV 或 JSON）。

2.数据分析与可视化：

使用 pandas 进行数据清洗与分组统计，用 matplotlib 绘制统计图表，如柱状图与饼图。

四、实验过程

（1）数据获取

通过爬取百度百科页面 “https://baike.baidu.com/item/披荆斩棘的哥哥”，解析其中参赛选手的基本信息。获取后保存为 list[dict] 格式，部分示例如下：

```
data = [
    {"name": "陈小春", "birth_day": "1967", "height": 173, "weight": 65},
    {"name": "张智霖", "birth_day": "1971", "height": 180, "weight": 70},
    {"name": "李承铉", "birth_day": "1984", "height": 182, "weight": 72},
    {"name": "胡海泉", "birth_day": "1975", "height": 176, "weight": 68},
```

{"name": "林志炫", "birth_day": "1966", "height": 178, "weight": 63}
]

（2）数据分析与可视化

① 年龄分布柱状图

以出生年份为横轴，人数为纵轴，统计参赛嘉宾的出生年份分布情况。

② 体重分布饼图

将选手按体重区间分组，展示比例关系。

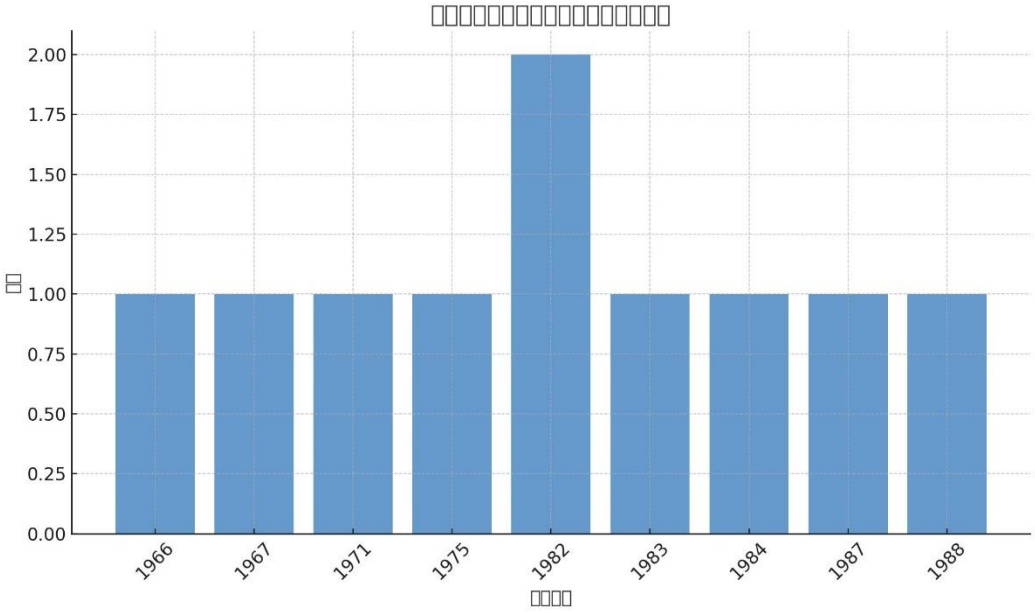
③ 身高分布饼图

统计选手的身高段落分布情况。

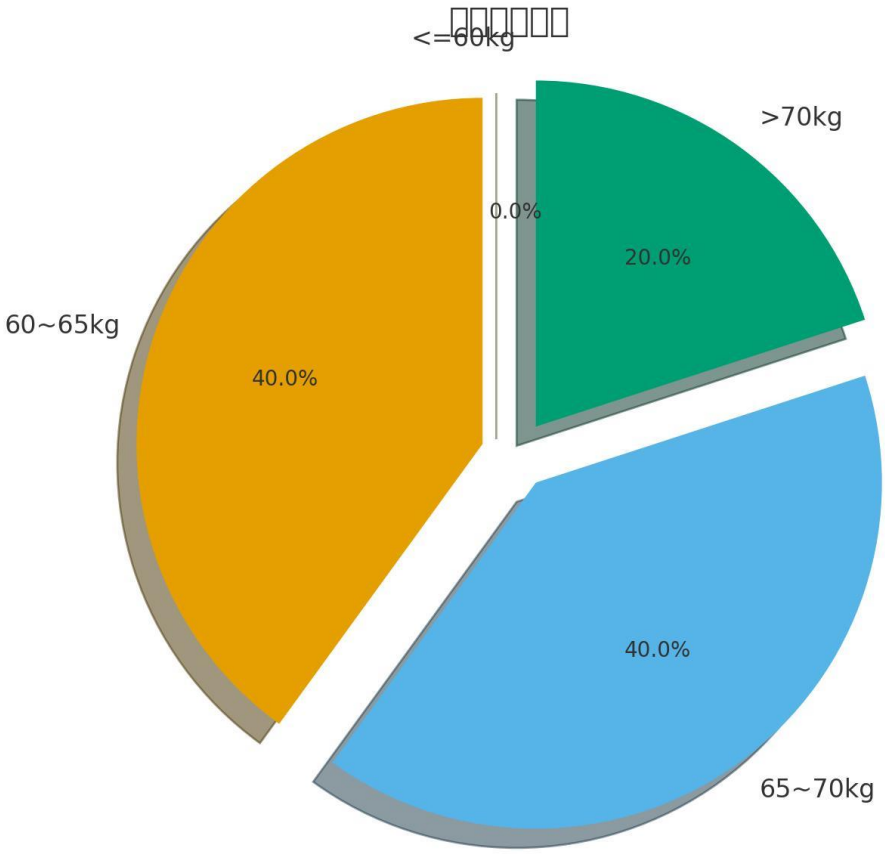
用飞桨 AI 进行实验具体图如下：

五、实验结果分析

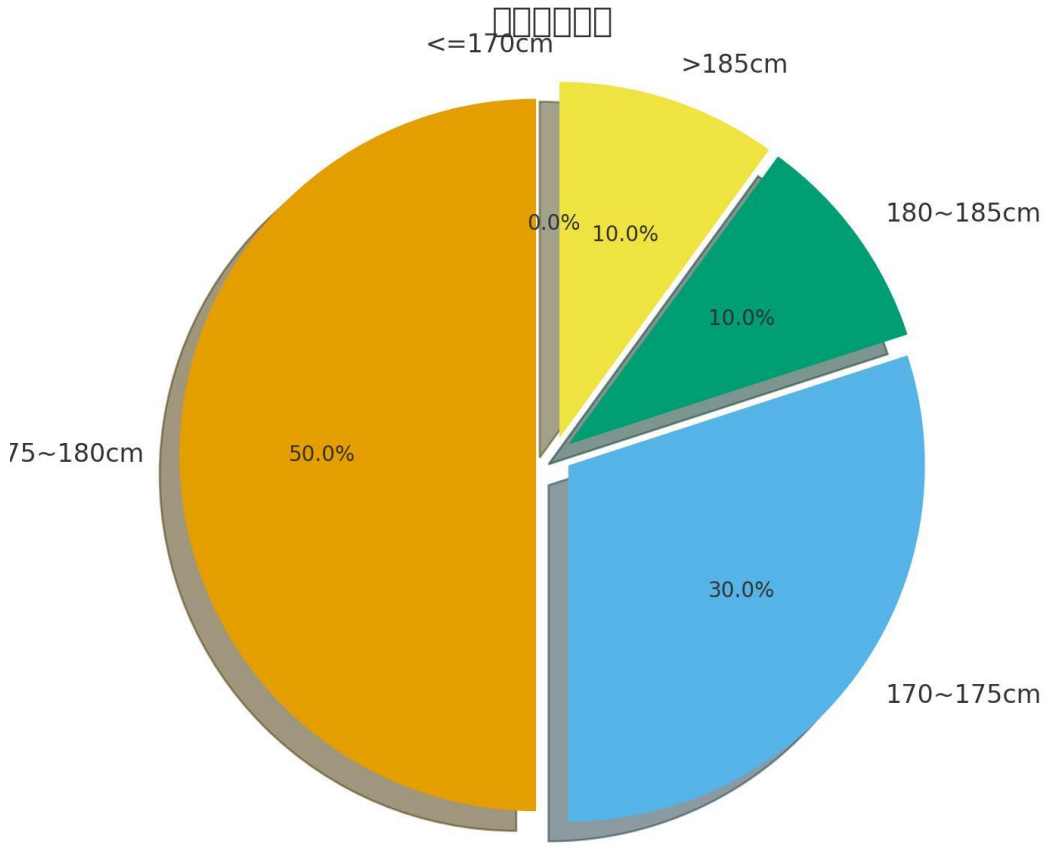
（1）年龄分布柱状图：



（2）选手体重分布饼状图：



（3）选手身高分布饼状图：



通过图表可以看出， 大部分选手出生在 1970－1985 年之间，说明节目的核心嘉宾多为 70、80 后。体重集中在 65－70kg 区间，占比最高。身高主要集中在 175－185cm 区间，整体形象较为均衡。

六、实验小结

本次实验成功实现了从网页中自动提取综艺节目选手信息的全过程，并基于采集数据进行了可视化统计。通过该实践，掌握了 Python 爬虫的基本流程；使用 pandas 进行数据清洗与分组；利用 matplotlib 绘制多类型可视化图表。

七、其它

通过此次实验，我更加深入地理解了“从网络获取到结构化数据”的全过程。实践中最

关键的部分在于：数据提取规则的确定；可视化表达方式的选择。这次实验为后续更复杂的数据分析任务打下了基础。