**大连理工大学《Python与数据分析》课程大作业**

课程名称： Python与数据分析 考试形式： 大作业

授课院 (系)： 软件学院 上交日期： 年 月 日

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 分值 | 作业一 | 作业二 | 作业三 | 作业四 | 总分 |
| 标准分 | 25 | 25 | 25 | 25 | 100 |
| 得 分 |  |  |  |  |  |

|  |  |  |
| --- | --- | --- |
| 姓名 | 学号 | 班级 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

一、作业一（共计25分）

得分

得分

**题目：**以胡润百富（https://www.hurun.net/）为数据源，完成以下任务。（1）爬取2024年胡润百富榜的榜单数据，包括财富值、个人信息以及企业信息。（2）对不同行业的富豪数量以及财富值进行统计，分析各行业的发展态势。（3）对爬取的数据，进行其他维度的分析，例如富豪在年龄、性别、出生地等方面的分布情况，利用柱状图、饼图、热力图等进行可视化展示。

**数据源**：胡润百富（https://www.hurun.net/）

二、作业二（共计25分）

得分

得分

**题目：**大连市2022-2024年天气数据分析与预测，任务描述：（1）爬取大连市近三年每天的天气状况（包含白天，夜晚） ，最高温度，最低温度，风力情况（包含白天，夜晚）（2）使用可视化工具（如Matplotlib、Seaborn、Echarts等）绘制近三年月平均气温变化图 ：计算每个月的平均最高温度和平均最低温度，绘制折线图展示一年中气温的变化趋势。（每个月取三年的平均值以下两题同理）（3）绘制近三年风力情况分布图 ：统计每个月不同风力等级（如 3-4 级、5-6 级等）出现的天数，绘制柱状图或饼图进行展示。（4）绘制近三年天气状况分布图 ：统计每个月中不同天气状况（如晴天、多云、阴天、雨天等）出现的天数，并绘制柱状图进行展示。（5）利用前三年的每月平均最高温度训练温度预测模型，给定月份生成对应月份的平均最高温度，并额外爬取2025年1-6月份数据，用折线图画出预测结果和真实结果曲线

**数据源：**天气后报https://www.tianqihoubao.com/ 。

三、作业三（共计25分）

得分

得分

**题目：**学术论文发表趋势分析 —— 基于DBLP的数据挖掘研究：选择计算机科学领域的国际顶级会议（例如 AAAI、IJCAI、CVPR、ICCV、NeurIPS，ICML，ICLR，KDD等），从 DBLP 学术数据库中爬取相关论文数据，进行如下分析任务：（1）选择3-5个代表性会议，爬取这些会议从2020年至今的所有论文信息，包括论文标题、作者、发表年份、会议名称、原始链接等。（2）对每届会议论文数量进行统计，绘制各会议的每届论文数量变化趋势图。（3）基于论文标题内容进行关键词提取，统计高频关键词，并可视化为词云图，判断2020年至今研究热点变化。（4）对已选会议在下一届可能发表的论文数量进行预测。

**数据源：**DBLP，谷歌学术，百度学术，国内镜像等。

四、作业四（共计25分）

得分

得分

**题目：**请以中国体彩网（https://www.zhcw.com）为数据源，完成以下任务：

1. 爬取截至2025年7月1日之前100期的大乐透开奖数据和中奖情况（链接：https://www.zhcw.com/kjxx/dlt/），分析大乐透总销售额随开奖日期的变化趋势并预测2025年7月1日之后最近一期的销售额。（2）根据爬取的大乐透数据对前区号码与后区号码进行频率统计与可视化，分析其历史分布规律。根据你发现的规律或者采用其他预测方法推荐一组大乐透号码，作为2025年7月1日之后最近一期的投注号码。（3）根据爬取的大乐透数据，分别统计周一、周三、周六的大乐透开奖号码和总销售额。对比不同开奖日之间的号码分布与销售额特征，分析是否存在显著差异或相似性。（4）爬取任意一个彩种中20位以上专家的公开数据，对专家的基本属性（彩龄、发文量等）和表现（中奖情况）进行统计分析，并通过可视化展示其分布规律、相互关系或对中奖率的影响。

**数据源：**中国体彩网（https://www.zhcw.com）

注：在报告中，需给出使用的重要库函数，贴出重要的代码（例如如何爬取、分析数据等），并给出可视化的截图。（红色字体提交时，请删掉）