**知乎用户数据分析**

**作品名称：**

**小组成员**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学号** | **姓名** | **院系** | **手机** | **邮箱** | **在项目小组中承担的任务（如编写\*\*功能部分代码）** |
| 191300020 | 黄彦骁 | 人工智能 | 13479572433 | AdrianHuang@smail.nju.edu.cn | 编写实验报告和ppt，编写爬虫部分代码 |
| 191300034 | 刘恬远 | 人工智能 | 18379052664 | 1064274091@qq.com | 编写数据分析，绘图可视化部分代码 |

**背景和问题**

|  |
| --- |
| 小组成员对于知乎用户相关的数据比如大V的粉丝数，知乎用户的学校，所在地等方面有很大的兴趣，因此萌生了做这样一个数据分析的想法  问题：1.爬虫时ip被封禁，而网上搜索到的免费ip质量不高，导致爬虫部分消耗时间过长  2.面对爬取到的数据，分布过于离散，导致很多图画出来既不好看也得不到有效信息，并且又不能放弃掉那部分离散数据  3.对数据进行处理时，有些方法并不完全适用，要进行多次尝试，选择最好的工具与模块 |

**设计思路**

|  |
| --- |
| 从张佳玮等十位大V的关注列表出发，”递归”式的爬取知乎用户数据10万左右，在去重，清洗掉异常数据之后利用数据可视化得出图像分析得出结论 |

**数据来源**

|  |
| --- |
| 知乎：<https://www.zhihu.com>  用户信息:网址过长，见数据源.md  关注列表:网址过长，见数据源.md |

**处理分析**

|  |
| --- |
| 主要利用pandas，numpy模块对excel文件进行分析与处理  并用pyecharts, matplotlib，wordcloud等工具进行可视化分析, 制作地图，词云，饼状图，条形图等 |

**结论**

|  |
| --- |
| 知乎男女比相对平衡  知乎用户分布沿海多，内陆少，且北上广地区密集  知乎用户所在大学（有记录）主要为985大学，所在专业与行业主要集中于互联网和计算机技术  大多数知乎用户回答、获赞、被喜欢的数量都比较少，但少部分用户这些方面十分多，说明知乎使用人群主要为一大部分普通人和一小部分大V  通过对前20名粉丝数量多的用户进行分析，发现粉丝数与获赞数、回答数、收藏数并不全是呈正相关，说明知乎用户对回答的质量要求高，只有真材实料才能被关注与点赞，表现总体比较理性 |

**其他想说的（可选）**

|  |
| --- |
| 我们对于网络爬取方面的能力还有所欠缺，希望以后学习到相关知识之后能对知乎做更加全面的分析  Pyecharts， seaborn等可视化工具十分美观，希望能多学习以便提供多种选择 |