10W+知乎用户数据可视化

选题汇报

李世林 2020210219

胡业媛 2020311222

彭晓 2020210941

背景

- 知乎平台是一个中文互联网的常用**知识问答**社区，其中用户的活跃程度分布、地理位置分布等等都值得进行**数据分析和可视化**。随着互联网社交的快速发展，出现了类似知乎治校的戏称，这反映了知乎正在成为**主流年轻人发表、获取知识的重要途径之一**。无论是提供舆情方向指导还是研究社群分布情况，知乎数据集都有价值。

数据集

- **10W** 份用户数据
- **200 MB**
- **18** 个基本属性
  - 7个数值属性
  - 3个文本属性
  - 1个地理属性
  - 4个子结构属性


```
{
    "answer_count": 0,
    "articles_count": 0,
    "avatar_url": "https://pic3.zhimg.com/v2-2f
    "business":
    {
        "id": "",
        "type": "topic",
        "url": "", "name": "",
        "avatar_url": "",
        "excerpt": "",
        "introduction": ""
    },
    "description": "北京斯坦威图书有限责任公司肇
    "educations": [],
    "employments": [],
    "favorited_count": 0,
    "follower_count": 11,
    "following_count": 1,
    "gender": -1,
    "headline": "用生活所感去读书，用读书所得去生
    "locations": [],
    "name": "斯坦威",
    "thanked_count": 0,
    "url_token": "si-tan-wei-51",
    "user_type": "organization",
    "voteup_count": 0
},
```


- `name`：昵称
- `description`：个人描述
- `educations`：教育背景
- `employments`：职业背景
- `gender`：性别
- `locations`：地区
- `answer_count`：回答数
- `articles_count`：文章数
- `voteup_count`：获赞数
- `thanked_count`：感谢数
- `favorited_count`：收藏数
- `follower_count`：被关注数
- `following_count`：关注数

预期设计-可视化方案

- 用户分布地理位置可视化
- 用户教育程度、职业背景可视化
- 用户社交统计可视化
- 用户描述词频统计
- ……

预期设计-全栈工作量分配

- **数据分析和处理**。（1周）
  - json解析，数据清洗工作……
- **前端设计**。（2周）
  - 界面GUI，哈希路由，响应式布局、动态效果……
- **数值数据的可视化**。（1周）
- **文本处理**。（1~2周）
  - 包括分词、去除停用词等。

已有结果展示


10W+知乎用户可视化
首页
简介
演示示例
10W+知乎用户数据可视化项目
小组成员
彭晓 计算机系
233
李世林 机械系
666
胡业媛 材料学院
233

在线演示