**开源软件基础大作业**

**全国高校数据统计与可视化分析**

**Data statistics and visual analysis of Chinese Universities**

学 生 姓 名： 林轩阳

学 号： 201892055

学 生 姓 名： 范子璇

学 号： 201892290

学 生 姓 名： 袁千禧

学 号： 201892256

完 成 日 期： 2021.1.5

开源软件基础大作业项目报告

目 录

1 [引 言 3](#_Toc7959)

[2 项目创新点 4](#_Toc14054)

[3 项目设计 4](#_Toc31349)

[4 项目实现 5](#_Toc1672)

[5 项目测试 6](#_Toc14074)

[6 项目分析 1](#_Toc23762)3

[7 项目不足和改进之处](#_Toc26162) 13

[8 项目总结 1](#_Toc14346)4

[9 成员感想 15](#_Toc7307)

# 引 言

开源软件基础是一门很全面的软件开发与测试方面课程，在这门课程中，我们不仅学习到了与开源软件项目有关的知识和理念，同时在课程的后半段学习了部分python语言知识和爬虫基础。在本次大作业中，我们从自身的角度出发，选取了“全国高校数据信息统计”这一主题，并通过各项技术对数据进行收集分类、可视化和分析，然后设计前端，集成了最终的数据网站。

本报告包括项目背景以及意义，项目创新点，项目设计，项目实现，项目测试，项目分析，项目不足以及可改进之处，项目总结和感想各个方面，对本次大作业进行一定程度的总结和汇报。

1 项目背景及意义

近几年来，国家政府和教育部加大高等教育体制改革、推动高水平院校建设，并在985、211计划的基础上提出了“双一流”院校规划，随着高校建设和高校规划的逐步推进，原本清晰的高校分布数据和高校数量统计又需要重新收集和统计。

在这种背景前提下，我们设计完成了本次大作业，本着对全国高校分布状况、高校水平和地区数量的统计分析原则，我们爬取生成了一系列柱状图、热力图和饼图，详细清晰地展示出了各地区高校数量、高校类型分布、各地区质量高校占比和数量统计等图表，便于需求者对高校分布数据进行一定的了解和查询。

# 2 项目创新点

为了避免数据罗列造成的印象模糊和理解偏差，通过图表的表现方式将高校数据信息进行全方位、多层次的展示。同时，我们在数据可视化的基础上，采用了饼状图、柱状图、热点图和水滴图等多种图表类型，通过各类图表的表现特性，例如饼图可以更好地体现出数量对比，热点图可以更好地体现出地区分布情况，而柱状图可以清晰地将数据进行从高到低的排序，如此一来，更加清晰地在不同图表下体现出差异和水平，让数据表现得更加详细突出。

# 3 项目设计

对于高考网的高校信息进行爬取，对数据进行处理与分析，针对高校的地区分布，类型，性质等做一些综合分析，并且将分析结果可视化，最后制作一个网站，这个网站有一个首页，它包括了这个项目的总体介绍，并且展示了我们的五个模块，点击进入不同的模块就可以看到对应的统计结果。

# 4 项目实现

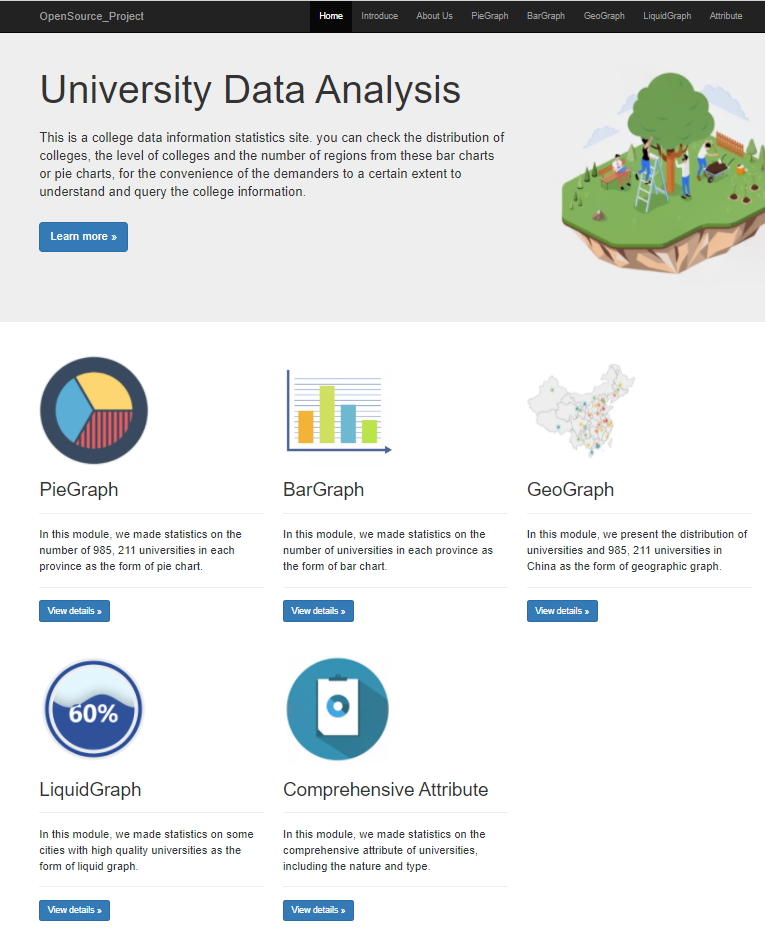
我们首先使用爬虫对于高考网的数据进行爬取并且将数据存入csv文件中。实现爬虫的程序是university\_crawl.py。我们要爬取的网页所有数据一共有107页，而且每一页的URL都是以 http://college.gaokao.com/schlist/p（）作为开头，而p后跟的数字就是页数。比如这个网站的第三页的URL就是“http://college.gaokao.com/schlist/p3”。利用这一特性我们设计了一个循环，循环次数即为页数，我们使用requests模块先将每一页的html进行爬取，然后用BeautifulSoup进行解析，用college\_list存放了每一页中的数据。然后我们设计了parse\_item函数，该函数能对每一行数据中的各个属性值赋到一个结果集当中并且返回该结果集。然后我们使用map函数将college\_list的数据给parse\_item函数进行处理，将处理后的数据存入college\_data.csv文件中。到此我们的数据爬取工作就结束了。

接下来就是用爬取到的数据做一些数据分析以及可视化了。实现这一部分的程序是analysis\_visualization.py。首先读取csv文件，然后使用pandas模块对于全部高校，211高校，985高校，高校性质，高校类型等数据分别进行了处理。然后用pyecharts模块分别制作了柱状图，饼状图，热力图，水球图，以及高校综合属性分析这五个模块，将同一个模块的图标放进同一个page当中，运行该文件后会生成5个html页面。

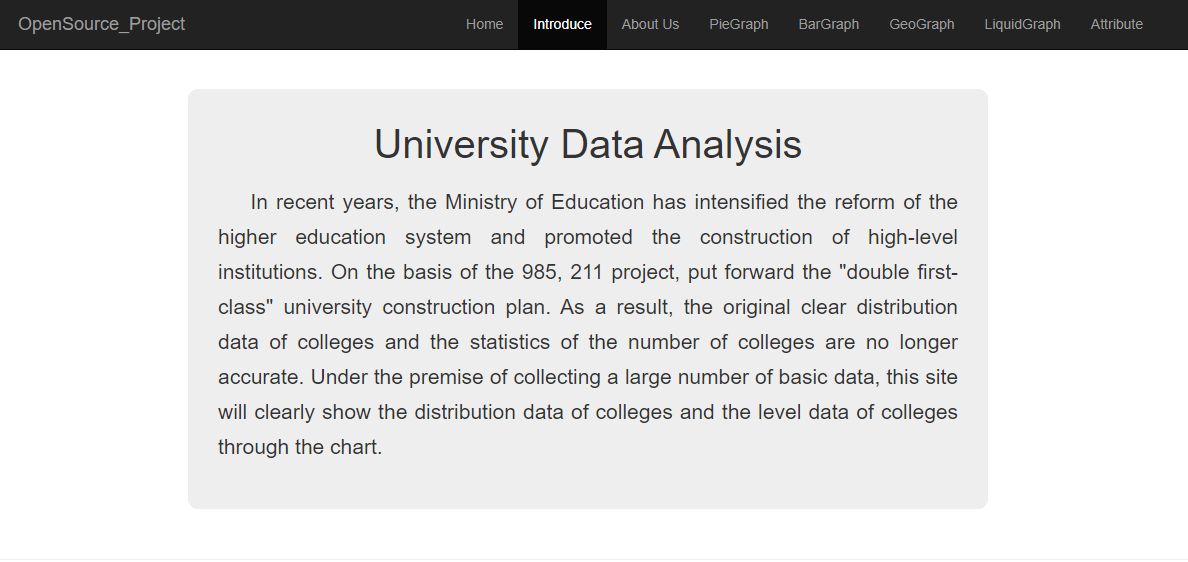
最后就是使用flask与bootstrap制作网站。首先在app.py程序中使用route函数将URL与调用的html文件进行了绑定。该网站使用了bootstrap提供的一个前端框架进行开发。这部分的主要工作就是对于网站首页与项目介绍页面进行了设计与制作，并且把pyecharts生成的html文件做了一些修改，使统计图表可以与框架进行适配。

# 5 项目测试

项目页面全局展示如下，整个界面的上方可以看见网站功能列表，网站功能背景介绍，以及项目小组人员信息。界面下方图形化的展示了主要功能，包括饼状图，柱状图，地理分布图，液体图，综合属性图。



主界面展示

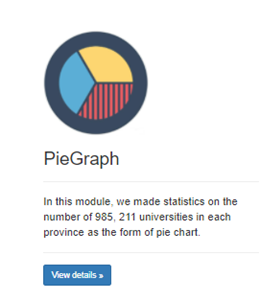


项目背景介绍

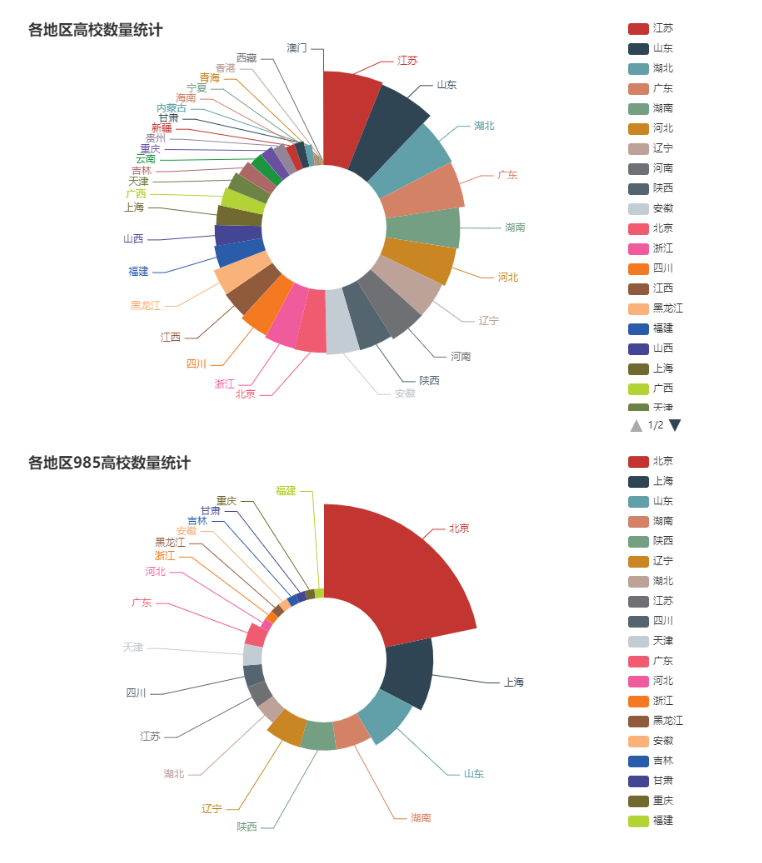


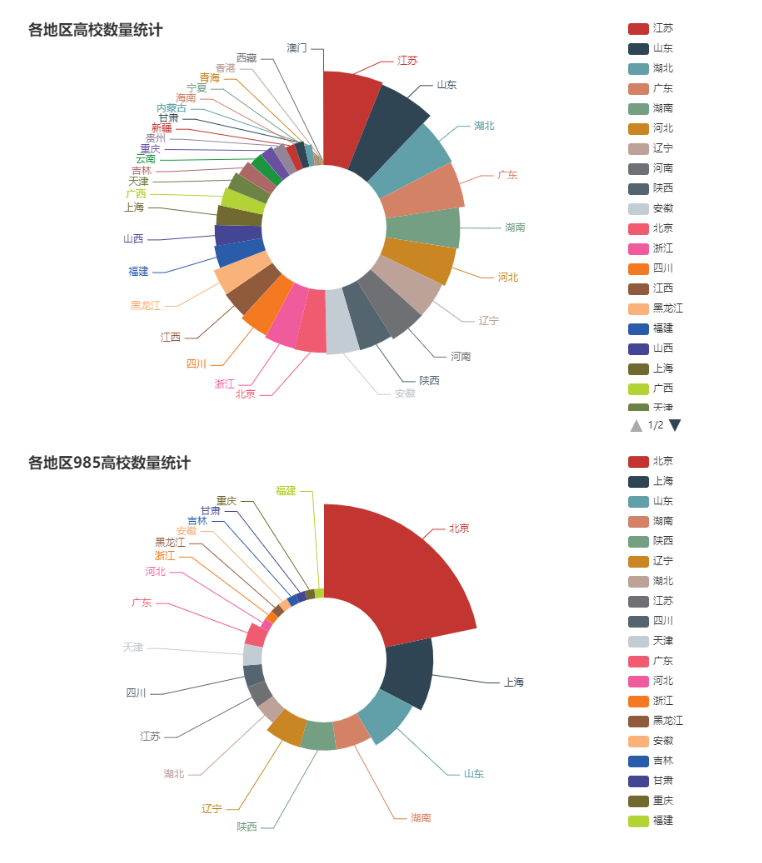
小组成员信息

5.1 饼状图界面：



饼状图界面入口

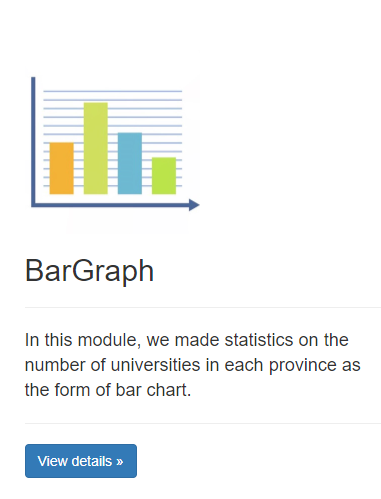




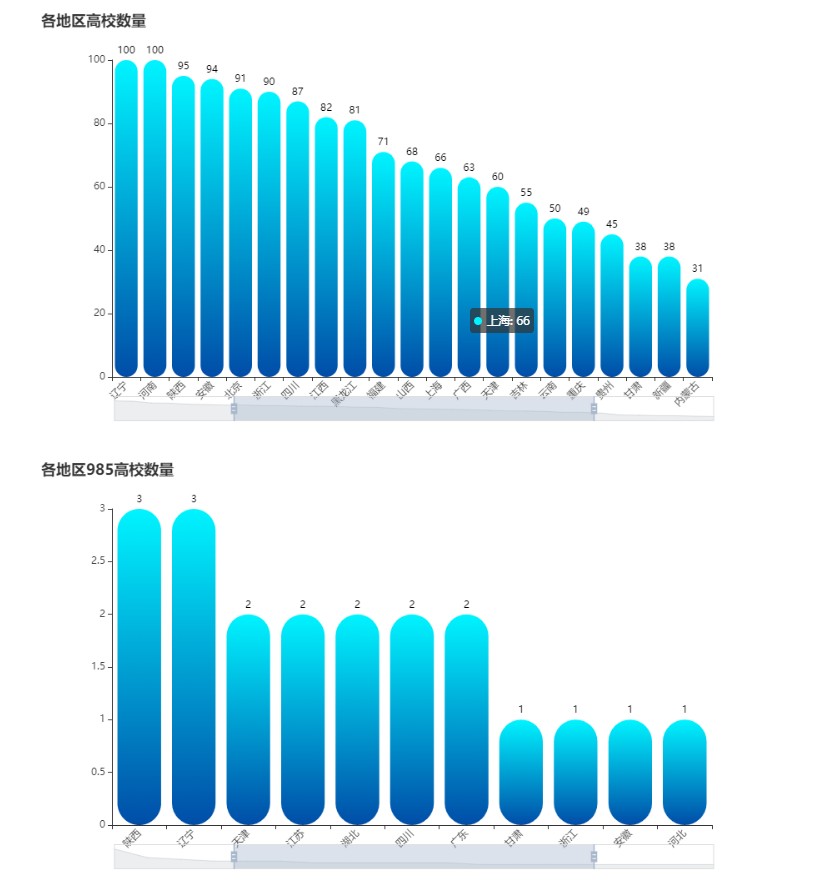
饼状图部分界面展示

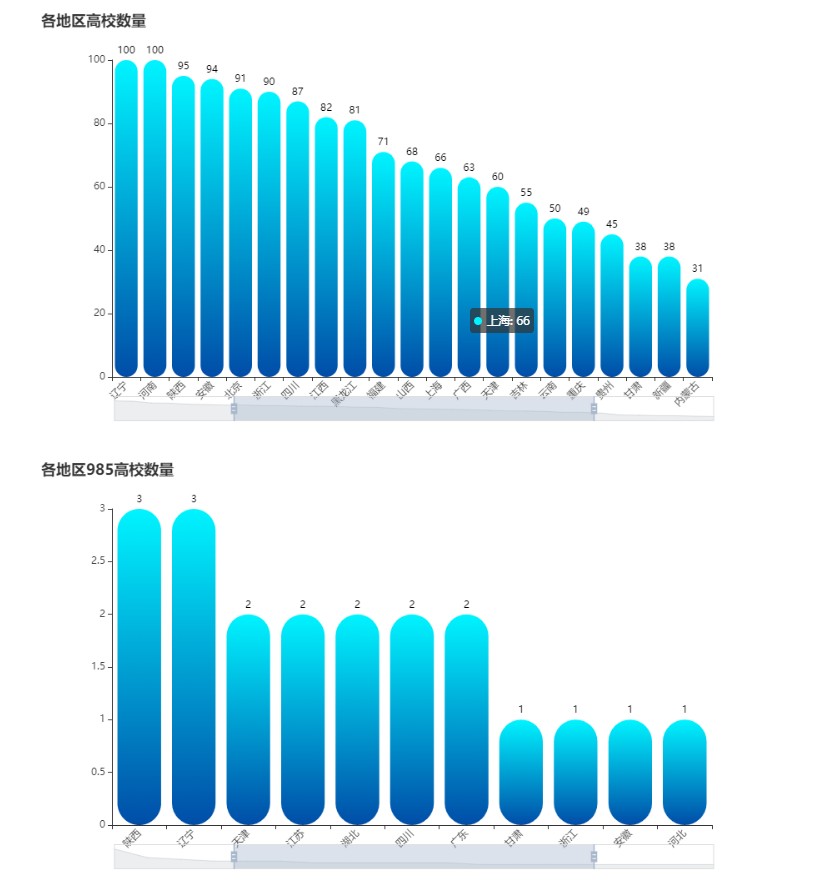
进入该界面我们可以在各个饼状信息图中看见各方面的高校信息统计如：各地区高校数量统计，各地区985数量统计，各地区211数量统计，全国各省高校数量前十统计，以及各省高校数量后十名统计。饼状信息统计可以让信息需求者快速看到关键信息如：北京在985，211高校的数量上占据显著优势；在高校数量方面，前十名并未形成显著差距，而后十名则差距较大，西藏地区在高校数量方面劣势显著，等等。将光标放置于图片的具体一项上可以显示具体数量，点击右侧的具体省份可以选择勾选省份是否参与统计，默认情况下所有省份都参与统计。饼状图的优势在于信息需求者能直观的看出各地区高校在各方面所占比重。

5.2 柱状图界面：



柱状图界面入口

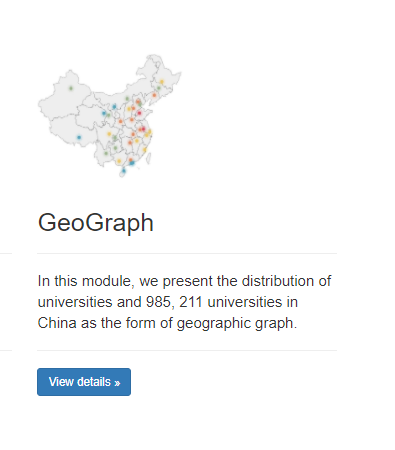




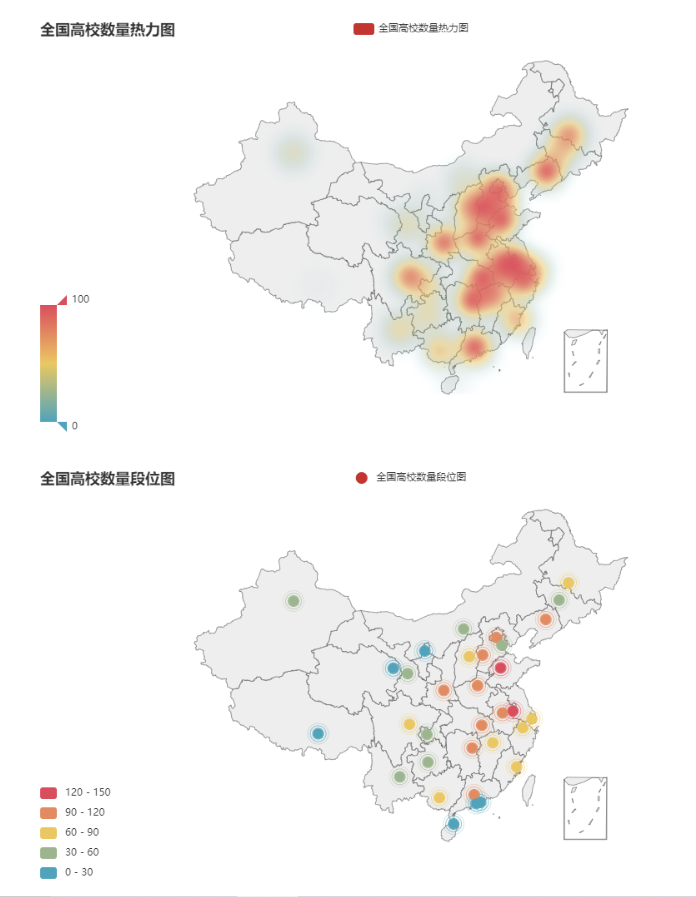
柱状图部分界面展示

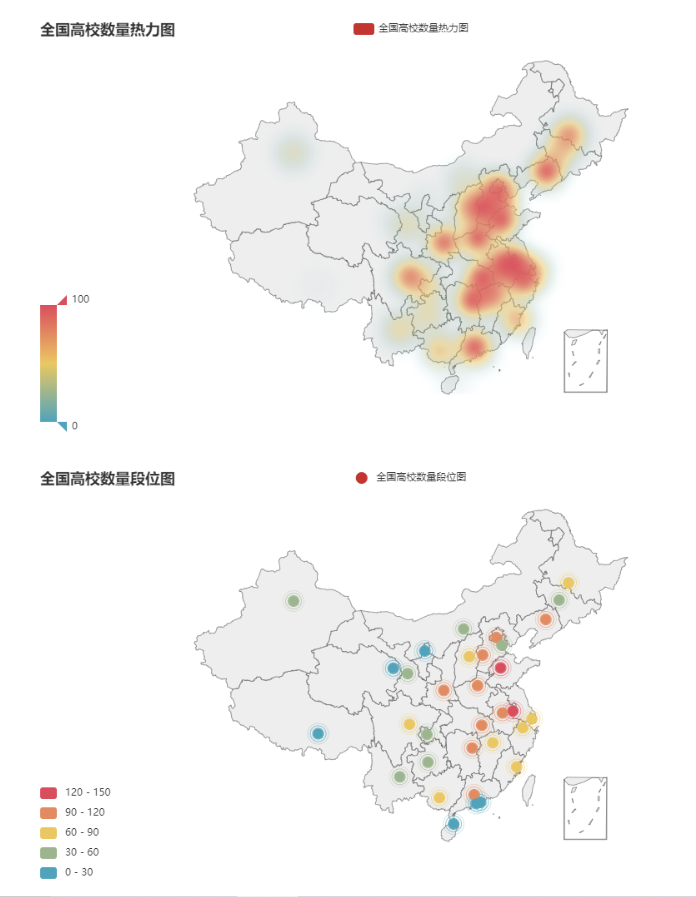
在该界面中我们可以得到关于各地区高校数量，各地区985数量，各地区211数量，以及各地区高质量高校数量的棒状信息展示。拖动滑块可以动态的清晰直观的看出各地区之间的数量差对比，数据的展示也更为直接明了。柱状图的优势在于信息需求者可以更加方便的将信息数量相近的地区进行对比。

5.3 地理分布图界面：



地理分布图界入口

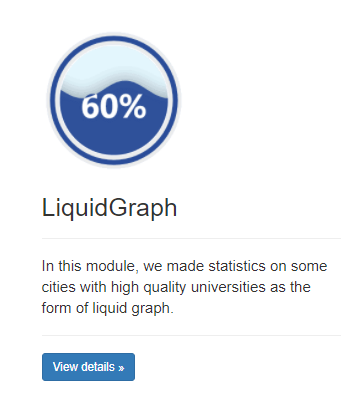




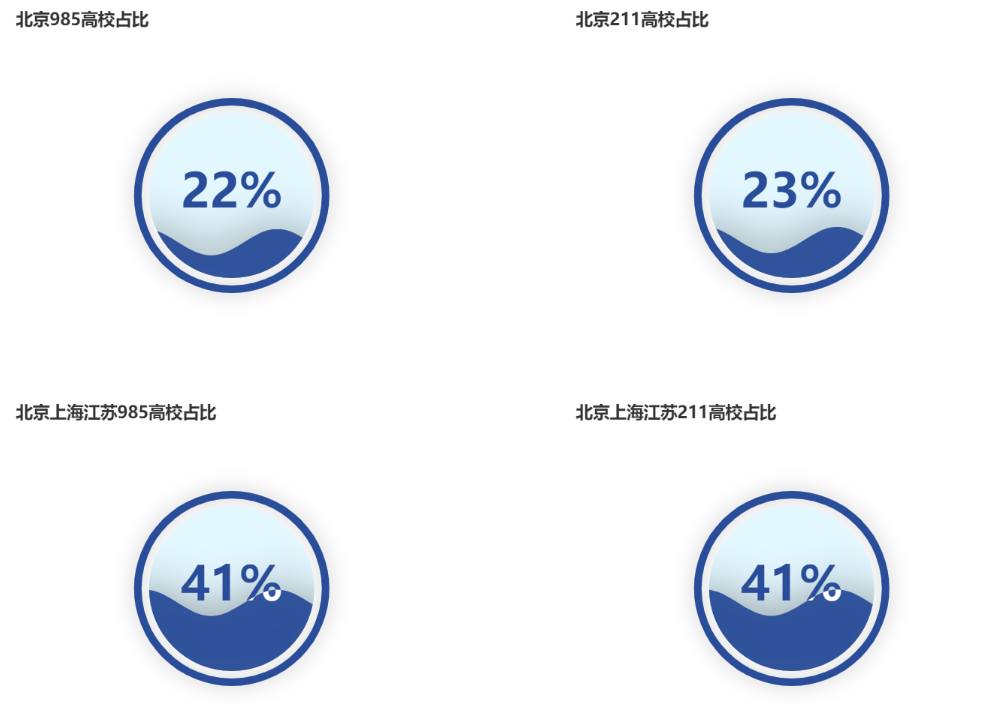
地理分布图部分界面展示

该界面展示了全国高校数量热力图，全国高校数量段位图，高质量高校分布热力图，高质量高校分布段位图，可以通过滑轮来进行放大缩小地图。该界面将高校信息与地理位置进行关联，使得信息需求者可以分析出高校分布与地理位置之间的关系，如：从全国高校热力图中可以看出高校在我国东部分布数量较多，密度较大等等。点击左侧的具体段位可以选择勾选省份是否显示，默认情况下所有段位会在地图中展示。

5.4 水球图界面：

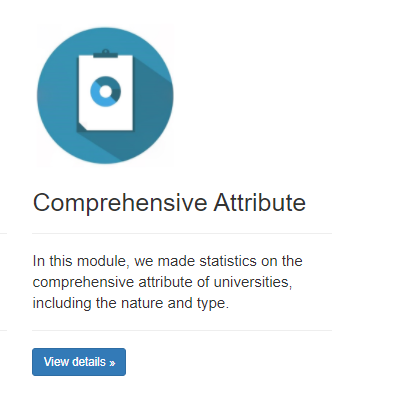


水球图界面入口

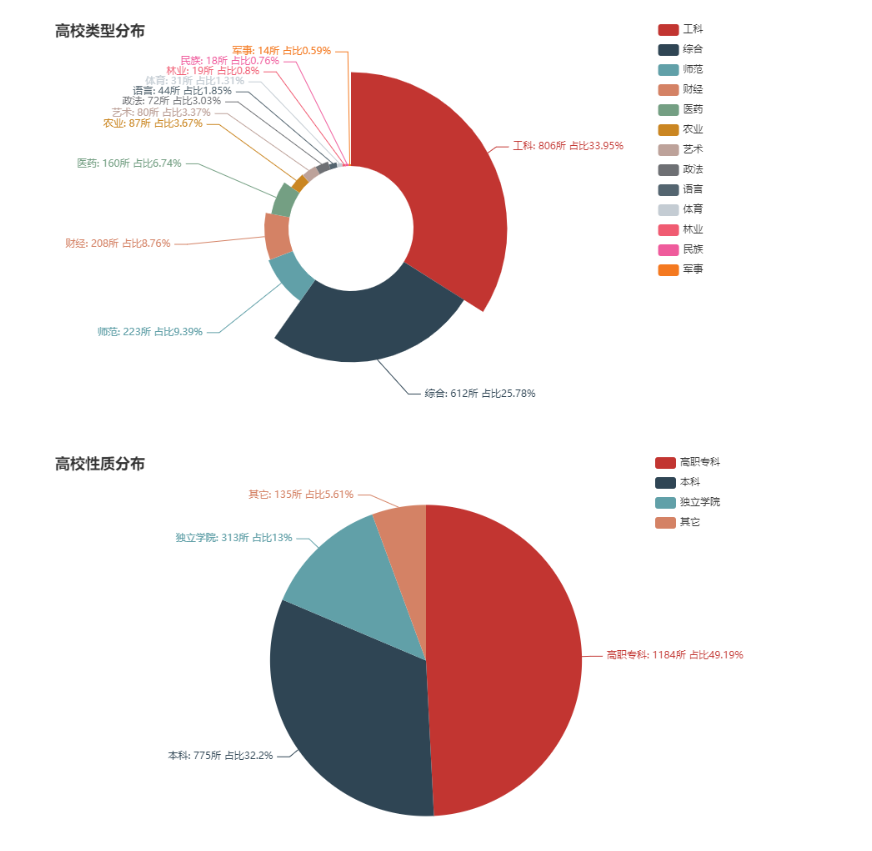


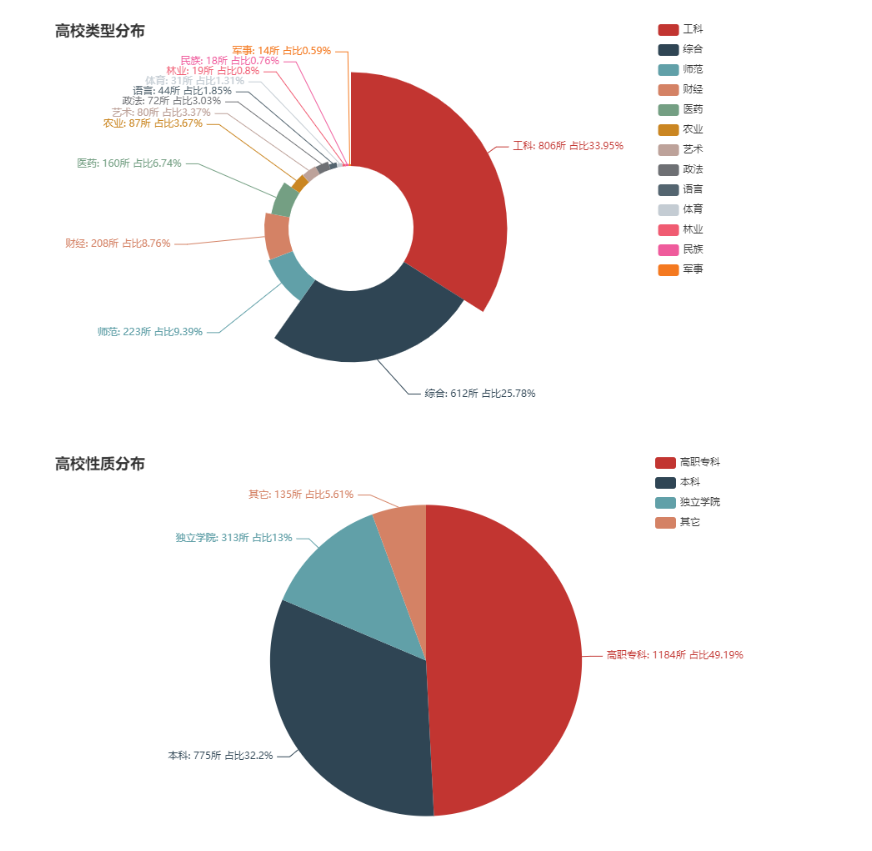
水球图界面展示

5.5 高校信息综合属性界面测试：



综合属性界面入口





综合属性界面展示

该界面用饼状图展示了全部高校的类型以及性质的分布占比，信息需求者可以得到相关信息如：工科院校数量一马当先，综合性大学则紧随其后，其他类型的大学百花齐放；而高职专科院校占据全部高校近一半数量，其次是本科占将近三分之一，等等。点击右侧的具体类型可以选择勾选省份是否参与统计，默认情况下所有类型都参与统计。

# 6 项目分析

教育事业一直是国民关注的热点，其中高校教育更是其中的重点话题。作为一名高校学生，这样的项目主体更加切合我们的现实，且该项目的内容可读性清晰，统计和分析结果便于展现，制表性良好，可行性高；在技术和经济支持方面，高考网和其他一部分高校官网公布了大量透明真实的高校数据和生源数据，很大程度上便利了我们对数据的爬取和统计；在社会需求度方面，近年的高校体制改革和建设计划使得一份全新精准的高校数据统计分析报告需求度逐渐提高，且项目偏差不会对使用者造成过大的负面影响。在风险因素方面，项目本身商用型价值低，数据统计较为粗略，仅供相关人士参考。

# 7 项目不足和改进之处

7.1 项目的不足之处：

(1)选择的网站所提供的高校信息量不够全面，缺少一些比较重要的信息，比如高校历年来的在各省的分数线以及省排名，高校的王牌专业等等，这也限制了我们对于高校做出更加全方位的分析。

(2)Introduce与about us界面做得有些简陋。

(3)由于爬取的网站中没有收录台湾省的大学相关信息，因此我们所做的统计中没有出现台湾省的大学的数据。我们在此郑重声明，从古至今，以至于未来，台湾省都是中国不可分割的一部分。

7.2 项目可改进之处：

(1)将Introduce与about us界面做进一步美化。

(2)可以多找几个网站进行数据收集，但是也会出现一些比较冷门的学校可能找不到信息的情况。

(3)制作更多种类的图表，利于使用者进行横向对比。

# 8 项目总结

8.1 成员分工：

林轩阳：撰写报告，选择目标网站，负责数据爬取，数据可视化的代码编写，设计与实现前端界面。

范子璇：撰写报告，负责URL与调用的html文件的绑定的代码编写，参与前端界面设计。

袁千禧：撰写报告，进行项目测试工作。

8.2 项目总结：

开源软件基础这门课程的课堂学习并不深入，更多的是要我们自己去探索学习，在完成本次大作业的过程中，我们不断摸索，不断学习，将未知的难题转化为已知的知识，很大程度的提升了自身。

这一次项目的开发我们运用了爬虫，数据可视化，前端开发等知识。爬虫的实现需要用到requests以及BeautifulSoup模块。这两个模块我们在课上也进行了学习，但是当时学的时候有些懵，这一次做项目的过程中又进一步地学习与巩固了这部分的知识，也深深体会到了Python功能的强大。那么下一步就是将爬取的数据做一些统计与分析并且可视化。这一部分我们使用的是pyecharts。Echarts是一个由百度开源的数据可视化工具，凭借着良好的交互性，精巧的图表设计，得到了众多开发者的认可。而pyecharts就是由echarts与python相结合而来的。这个模块功能十分齐全而且方便，我们只需要对图表的options进行配置之后，放进对应的page中，再使用render函数对page进行渲染，就会自动给我们生成一个html文件。但是光有html文件就太简陋了，于是我们就使用flask框架制作了一个网站。我们使用的前端框架是bootstrap官网提供的，不过网页的布局以及图片都是我们自己进行设计的。

总的来说，这一次的项目开发不但让我们巩固了一些课上学到的知识，而且还促使我们去网上学习了一些新的知识，让我们受益匪浅。

# 9 成员感想

林轩阳：

这学期我们学习了开源软件基础这门课程。在这门课上，我们学习了一些python基础以及一些python的库，我觉得十分有意思。在这学期末我们小组运用这些python知识做了一个项目——全国大学数据分析。在做项目的过程中我收获颇丰。例如实现爬虫需要用到requests以及BeautifulSoup模块，但是当时在课上学习这部分知识时有些懵，通过这一次做项目进一步地巩固了这一块知识。我还学习了pyecharts这个库。我只需要把图标的相关配置写好，执行代码之后就可以生成一个html文件，真的是十分方便。然后这一次的前端界面也是由我设计的，感觉做前端界面还真不是那么容易的。找图片，图片尺寸的统一，界面的排版，间隔什么的都是需要考虑的。不过当自己把这个网页做出来的时候还是很开心的。通过这次的项目开发，我体会到了python的强大，数据统计的深奥，以及前端开发的魅力，以后的日子我也会继续学习，脚踏实地，不断充实自己。

范子璇：

本次开源软件基础课程大作业的制作，我们在开源软件课程的学习基础之上，继续学习python基础知识、爬虫相关技术并自学网站框架的架构和制作，在积极巩固老师讲授课程内容的基础上，拓宽自身视角，为大作业的技术方面打下更坚实的基础。项目从选题开始，我们便从实际出发，选取较为切合自身的“高校数据分析”主题项目，并且通过各大网站进行数据的爬取和统计，并对得到的内容进行整合填充和打磨，并在项目报告编写阶段不断发现错误，改进缺陷，得到最终的项目成果。不过，在本次大作业的制作中，我的技术贡献太少，代码编写和网站设计大多由组长完成，但在大作业的完成过程中，我也积极学习相关技术的内容和方法，和其他组员一起进步成长，在以后的项目制作中，我一定会更加努力地自学相关知识，争取在核心技术方面做出更多贡献。

袁千禧：

本次开源软件课程我们学习了很多知识，也加深了我对程序设计的理解，一定程度上提高了我的编程能力。本次的大作业内容对我们来说是一次挑战，我们需要用爬虫从网站是爬出数据并将获得的数据进行处理，用一个可视化的前端将数据分门别类进行处理展示。在本次大作业中我主要负责项目测试，我们的项目有关于高校数据，在测试过程中我也对中国高校的分布等信息有了进一步的了解。在大作业的进行当中我明白了团结就是力量，任务的成功进行需要大家相互交流协作，才可以让最终的成果展现。软件工程专业的学习是一点一滴逐渐积累而成的，正所谓，不积跬步，无以至千里，不积小流，无以成江海。希望在以后的学习过程中，我可以脚踏实地一步步前进，有所收获。

参考代码：bootstrap的前端框架：https://v3.bootcss.com/getting-started/#examples