**开源软件基础大作业**

**爬虫与数据可视化**

|  |  |
| --- | --- |
| 系别 | 软件工程 |
| 学生 | 朱金龙 |
| 学号 | 201992291 |
| 班级 | 软测1905 |
| 成绩 |  |

1. **项目背景**

**众所周知github是一个全球最大的代码交流网站，当然我们国内也有一些不错的技术交流分享的知识网站，例如csdn ,博客园等等。那么本次大作业主要是针对于csdn中的开源模块中的大家上传的项目文件的数据信息进行爬取，并做出一些相关的数据可视化分析等等。本次大作业中主要使用到5个关于python的开源模块，它们分别是requests , re , pandas ,pymongo ,pyecharts。**

1. **项目设计**

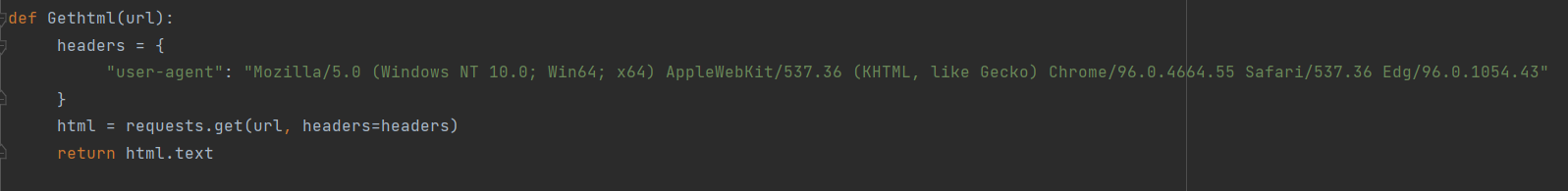
**本次大作业主要分为3个阶段:**

**第一阶段：**

**通过 requests进行网页html文件的获取，然后对html中的关键信息进行re正则提取**

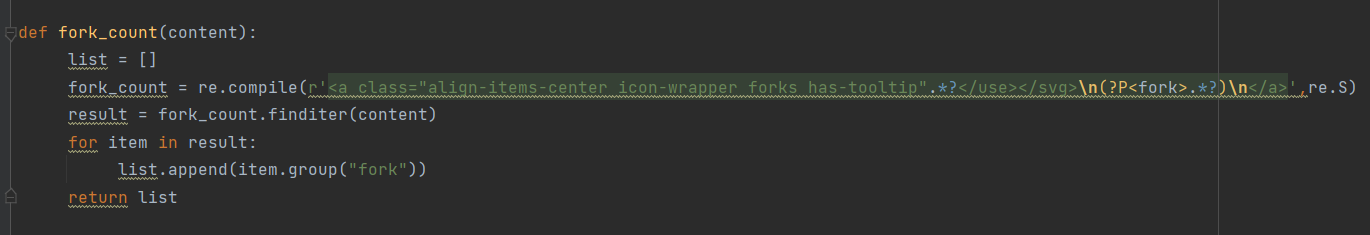
**具体代码与运行截图：**

1. **requests模块代码展示：**

****

1. **re模块代码示例：**

**这里我们一共提取了页面上的8种信息，分别是：name, project, describe, star, fork, 合并请求，issue ,license这8种信息。每个正则方法提取方法大致如下：**

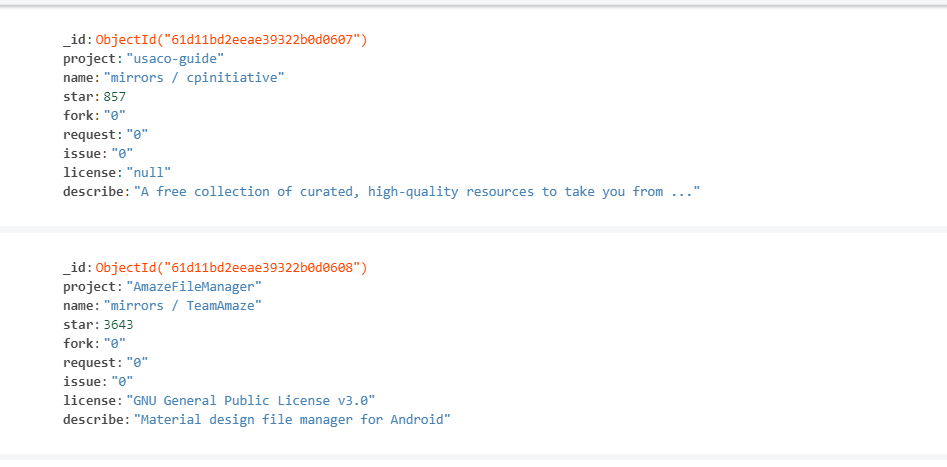
****

**第二阶段：**

**实现1步骤后，对多个网页上的信息的提取，并将数据存入本地mongodb中**

**具体代码与运行截图：**

**这里我们循环访问目标网页，通过每次更改url最后的数字，来实现下页的html的获取，对于有些不存在的信息元素我们暂时假定其为null,然后写入数据库：数据库中数据大致如下所示：**

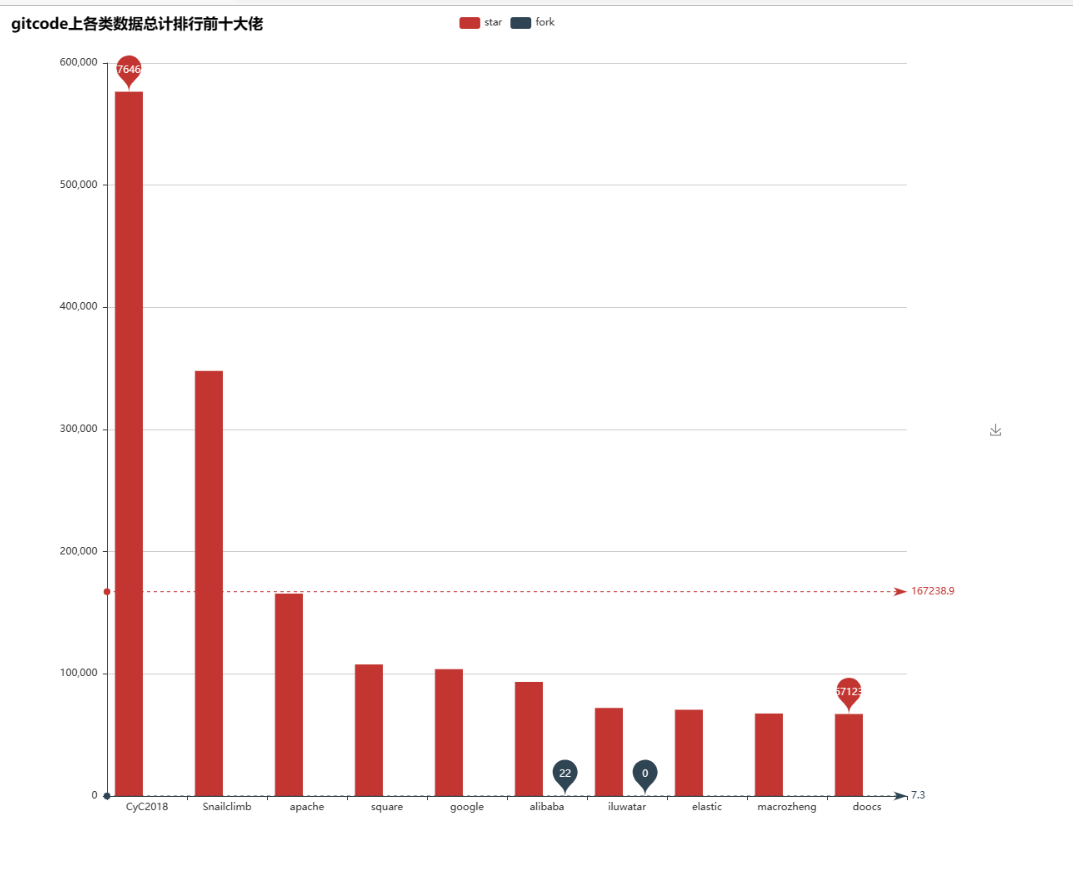
****

**第三阶段：**

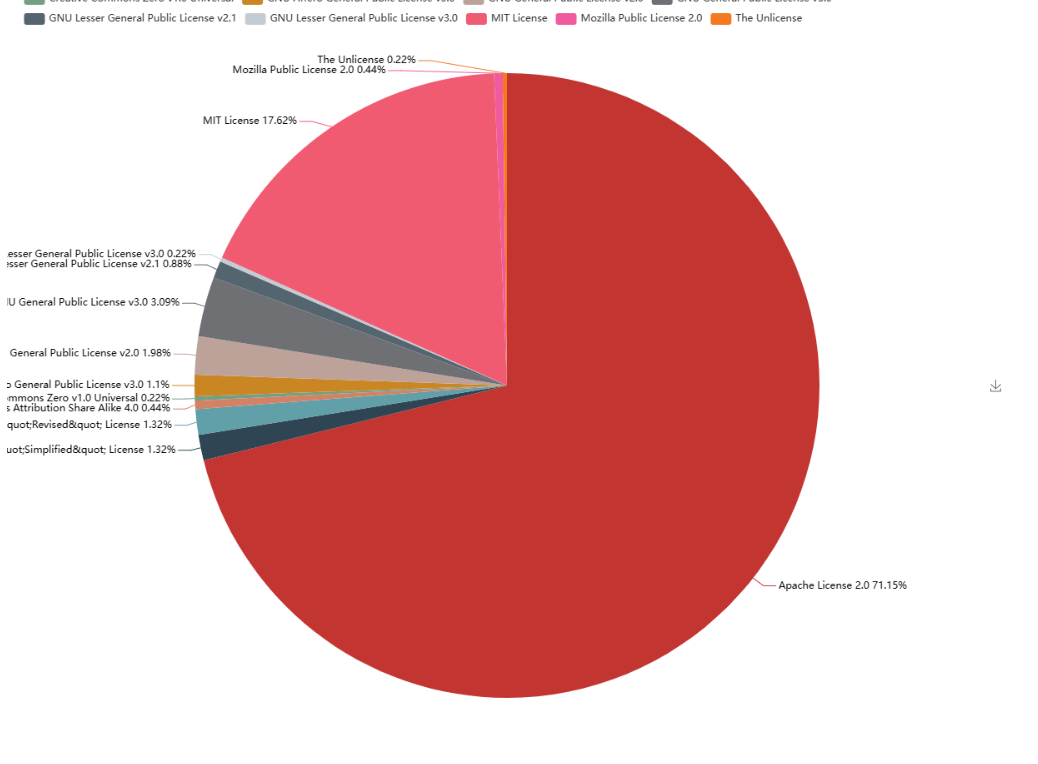
**这一步主要是对存储的信息进行数据分析，这里使用到了pandas这一模块的功能，我们首先将数据库中的数据转换成pandas中的dataframe模式，接着对我们想要的信息进行提取，聚合，排序等等操作，然后将我们筛选过后的信息，转换成列表类型，然后通过pyecharts的强大功能，生成2 个html文件作为我们最后的成果展示。**

**具体代码和运行截图：**

1. **pandas的数据分析，通过观察项目我发现在众多开源项目中，并不是我最开始所想的每个项目对应一位作者，而且有很多由一位作者创作的项目可能是一个系列，以至于他们的名字相近，所以我们可能不能直观的看出一个项目的优秀之处，例如一位名为cyc2018的作者，就写了很多一样名称的项目，甚至简介都没有改变，这对于可视化来讲是没有直观性的，所以在这里我决定先针对于这些已有项目的作者进行排名，通过获取的star数目，来可视化出排名前十的大牛。**

****

1. **接着我有发现每个作者都喜欢对于自己的项目设置不同的license,于是我再次统计了项目中众多开源协议的使用占比，并用pyecharts中的Pie图实现了它。**



1. **项目分析**

**本次作业中的不足主要有一下几点： 1，数据量太少，由此可以看出，在国内呢的开源项目是尚未开发的领域，我们也需要努力的大造出更好的开源社区环境。另外一个问题是本次作业不能实现很好的移植性，要求运行环境必须在本地部署mongoDB，针对于如何改进，则是生成csv文件，即将数据全部存入.csv文件，在pandas调用时也是只需要读取csv文件就好。**

**在测试中我也发现了这样的一个问题，我最初爬取数据的时候，是将数据存在数组中的，当时是724个项目，共37页，但是在我第二天继续运行时却发现之前的代码无法运行，始终提示数组越界，我们可以看到这个项目的个数是在不断增长的，所以我们需要为为程序安排适当的空间，做出一些预防性维护。**

****

1. **项目总结与感想**

**这次作业主要是进行了一次基础的爬虫和数据分析与可视化，在编写的同时，也感受到了，python语言的便捷性，甚至可以用1行代码来代替java甚至是c这类语言的多行代码，众多已有的开源模块为我们提供了更加丰富的功能，不需要反复的造轮子。这种开源的环境真的是让大家享受更好的编程体验。**