**Python大作业设计方案**

**班级：电信科20-1班 姓名：洪祎麟 学号：2020217793**

在本次Python大作业中，我主要是针对四个问题进行了分析与研究。分别是：

1. 动画片在近年来的发展现状。
2. 了解动画片具体讲述的主题，统计动画片在当季播出之后的观看情况
3. 根据一部看过的动画片，分析与该动画片最相似的动画片。
4. 根据用户所输入的动画片，搜索出对应动画片的相关信息

对于这四个问题，我大体上均采用了从网站上爬取数据->对数据进行清洗处理->对清洗后的数据进行可视化分析这样的方式进行解决。我选择用来获取的网站是国内最为权威的动画评分平台之一：Bangumi，在Bangumi上获得的数据相比其他平台更为真实准确，适合用来进行数据分析。针对本题的可视化需求，我使用了Pyecharts库，用于v1版本的使用较为复杂，故选择了0.5.11版本对数据进行可视化。对于数据的爬取上，我使用了requests库进行爬取，并采用re库和bs4库对所需要的数据进行提取。在爬取时我使用了自己电脑所在浏览器的header进行模拟，以便于爬取能够准确进行。对于数据的集中处理统计，我是用的是pandas库，进行一系列操作。针对评论区中部分出现乱码的问题，我采用了xlsxwriter库中的方法，将评论区顺利保存入xlsx文件中。对于具体动画评论区的评论分析，我采用了jieba,wordcloud,matplotlib和imageio库，对评论区的评论先进行中文的分词，并导入停用字表，删除评论中一些错误的分词，使得分词结果更加精确。对于每一部分所爬取的数据，我均已xlsx格式存储在本地，每一部分可视化操作得到的图表也均已html格式存储在本地。

对于最终结果的展示，我使用了Pyqt5这一GUI库制作出GUI，来展示每一题最终数据分析后得到的结果。下面对具体每一个问题的处理方法分别进行介绍：



针对第一问：对于动画片近年来的发展现状，我通过爬取Bangumi动画排行榜的前100页，2400条数据来分析大题情况。分析的主要数据如下图：



针对第二问，对于获得动画片的观看情况，我主要选取了两部动画《CLAS》和《摇曳露营》为例，通过对动画评论区进行提取分析，制作相应的词云数据进行分析。

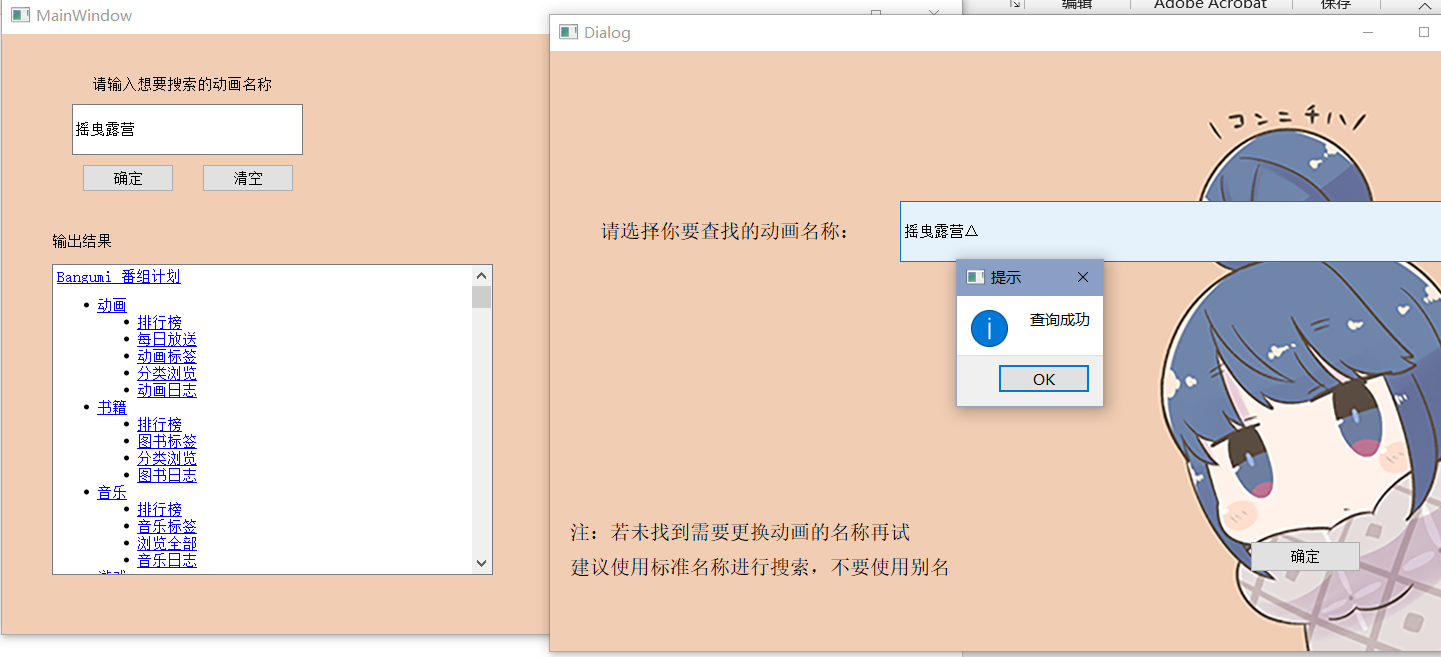


针对第三问，对于分析一部动画片所类似的动画片，我通过分别对动画片的标签所对应的前一页动画片，和评论区中的用户所最喜欢看的前10部动画片进行分析，得到看过一部动画片的人更适合看的其他动画片。



对于第四问，我通过gui的交互页面，由用户来输入要查找信息的动画片名称，从对应网站的搜索框对应选择具体的动画片项，最终确认，即可在窗口中看到对应动画的html。





本次大作业最终所生成的数据文件如下：

