**第一次设计报告: 夏远林 PB19020632**

**概述：**

我本身是统计专业的，所以GUI的实现仅仅是结果展示，更多的功能和具体的方法实现都在Analysis.ipynb以及datagetter.py当中，故而整体展示会略显单调，但是会详细说明具体的图片以及数据的由来。

**数据获取部分:**

爬取的豆瓣Top250的电影数据进行粗略分析，电影的信息存储在data/DriginalData.josn当中，主要分析描述的是电影的:

1.评分以及评价人数  
 2.电影的制片国家/语言/上映时期(几几年上映)/片长

3.电影的类型(剧情还是动作还是爱情etc.)

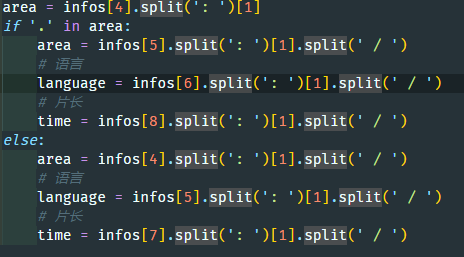
数据获取通过的是request进行源信息获取，即获取对应的html，然后通过Beautifulsoup进行筛选提取，最后存储到json文件当中.

**函数功能以及说明:**

首先因为豆瓣top250是分成10页的，也就是每页导航25个电影，所以第一步需要获取的就是根据每页的导航获取具体电影的链接，从而通过这个链接访问到具体的电影界面，获取数据，那么第一步就是通过getonepagelist实现，具体可参考datagetter.py

随后因为获取到了具体的电影链接，下一步就是申请访问这个链接并且通过Beatifulsoup进行分析，主要通过getfilminfo实现。(因为豆瓣的反爬问题，所以每次爬下来一个电影就需要进行一次数据的初步清洗以及存储，通过正则表达式以及json标准库自带的函数实现。)

**(数据获取部分的碎碎念):**

 爬数据真的烦，每个网站几乎都有反爬，既要伪装成浏览器，又要反复切换代理ip，我有点懒所以是手动切换代理ip，如果爬取的数据的数量级上升一个或几个，我或许需要制作一个ip库进行随机抽样来规避风险了。而且在html当中电影的数据结构竟然是没有index的顺序数组，实在是太愚蠢了，搞的有些电影的上映时间的信息在a[6]有的在a[8]

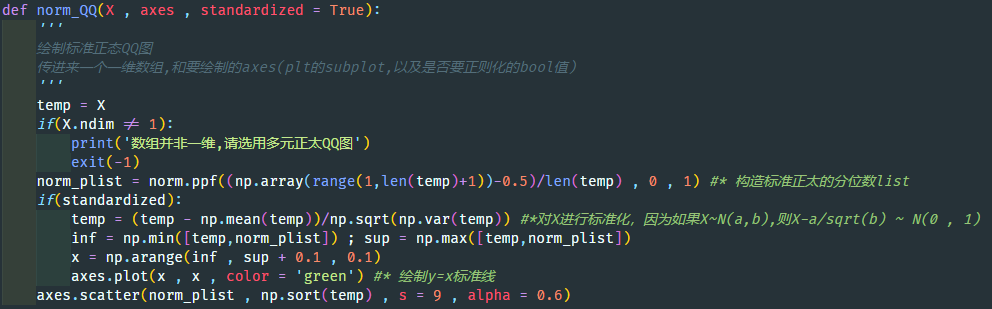
**数据分析部分:**

数据分析主要通过Analysis.ipynb实现，主要模块包括:

1. 数据的读取和分类
2. 数据的描述性统计(直方图，散点图)
3. 数据的粗浅分析(只做了一个正太分布的qq图，其实可以更进一步做一下回归和模型建立)

**非标准库以及第三方库的函数说明:**

**** 可以从这里看到，我引用了一些奇怪的函数，这些函数的来源是我放的第二个连接<https://github.com/CiscoXY/OwnFunction.git>

 本次作业主要用的函数就是static\_plt.py当中的norm\_QQ，即:

norm\_QQ的形参分别为:一个numpy的一维array,一个matplotlib的axes,来源于’fig , axes = plt.subplot(………)’，绘图原理是如果一个数据服从正态分布N(a , b)

那么对这组数据进行标准化后，数据应该服从标准正态分布，所以分位数也应当是标准正态分布，即将标准化的数据的分位数与标准正态分布的分位数做scatter，从而对比是否是一条y=x的直线，越接近直线越说明服从正态分布，反之就不可信。函数中norm\_plist就是构造标准正态分位数的数组，其中norm.ppf是scipy.static中的函数，主要功能是提供均值和方差，以及所要获取的右侧百分位数的值(可以是数组，也就是那个np.array)，从而得到分位数

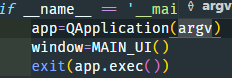
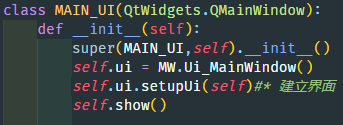
其他数据的描述性和分析工作都可以详见Analysis.py文件当中，需要注意的是我特地设置了每个图都是(9,9) dpi都是110，从而形成990x990像素的图片，方便嵌入pyqt5的label当中进行展示。

**(数据分析部分的碎碎念):**

我感觉pandas和seaborn的接口做的有点差，可塑性有丶低，虽然集成度和调取很方便容易，但是很不容易修改，应该提供一个接口，可以让我主动选取放在哪个位置(就像我自己的norm\_QQ一样)，这样我可以实现很简单的一个fig当中放一大堆子图，而不是每次绘图都得新建一个fig，并且根据pandas和seaborn的绘图进行略有麻烦的调整。

**那么最后的最后就是结果的展示了，通过pyqt5的GUI窗口实现，主要相关的文件是UI\_packages当中通过designer转换而来的Ui\_MainWindow.py以及主文件当中的main.py文件。**

**结果展示部分:**

**** 在main.py当中定义了继承的新类：class Main\_UI 。其中通过super继承了QmainWindow这个类，并且对self新建了一个属性.ui来链接设计出来的Ui\_MainWindow窗口。随后通过后面的信号与槽链接button和方法，方法主要就是控制label显示对应的图片。其中图片显示通过PyQt5.QtGui.QPixmap实现。

最后在结尾部分调用即可。