微博爬虫数据分析可视化程序设计报告

# 1 需求分析

## 1.1 引言

大数据时代的到来，让数据成为商品，数据分析也成为一个炙手可热的领域。可视化作为数据分析越来越重要的一部分，不仅让数据分析结果更加直观，还能发掘更多有价值的信息。微博等一些社交数据的挖掘与分析是近年来的热点，本文的程序设计对现有的微博爬虫[1]进行改编，用Python3重新编写原有的Python2的代码，并借助PyQt5(插件版，非软件)工具编写可视化界面，最后运行得到可视分析结果。

## 1.2 功能需求

1.界面：

输入用户（爬取对象）微博ID（如图1所示，网址中的数字）、输入本机Cookie（如图2所示）、数据分析结果显示



图1 用户ID

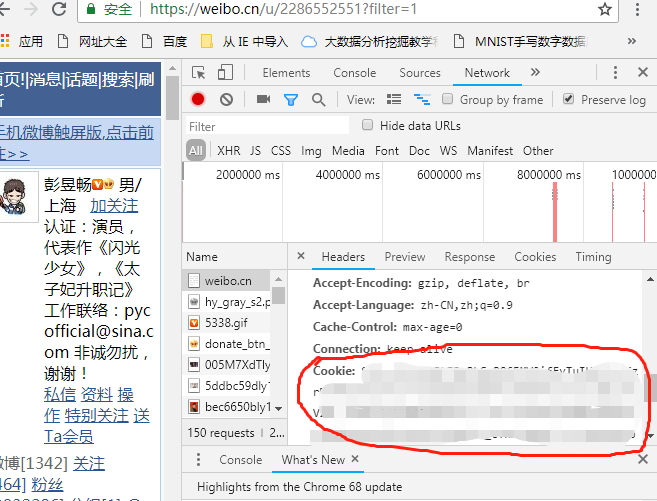


图2 Cookie获取

2.微博数据爬取：

根据界面输入的用户ID和Cookie爬取该用户的微博数据（原创微博文本）。

3.数据分析并显示：

根据每条微博中的关键字对微博进行分类，如美食类、旅游类等，结果进行统计并展示。

## 1.3运行需求

运行环境：Windows 8

开发环境：Eclipse、Python3、PyQt5

# 2 详细设计

## 2.1界面设计

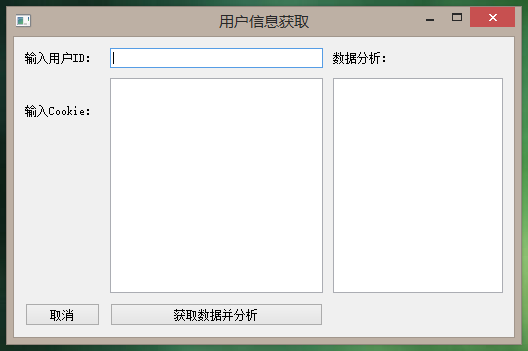


图3 程序界面

## 2.2 程序设计

如图4所示，程序包括4部分：界面(UI.py)、微博爬虫(WeiboCrawler.py)、数据分析(analysis.py)和关键字字典(keywords\_new.py)。

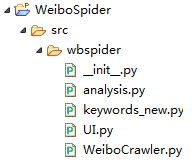


图4 程序目录

界面用Python中的插件PyQt5编写，当输入完用户ID和Cookie后，点击“获取数据并分析”按钮（如图5所示），爬虫函数获得userid和cookie两个值（如图6所示），并开始爬取数据（控制台显示爬取的数据，如图7所示）。



图5 输入信息

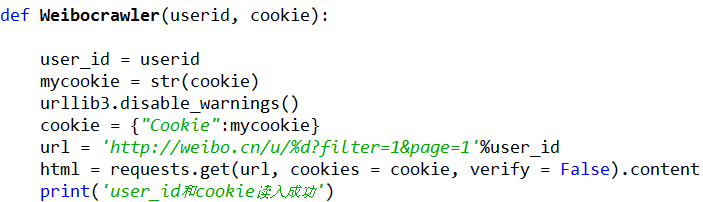


图6 获取两个值

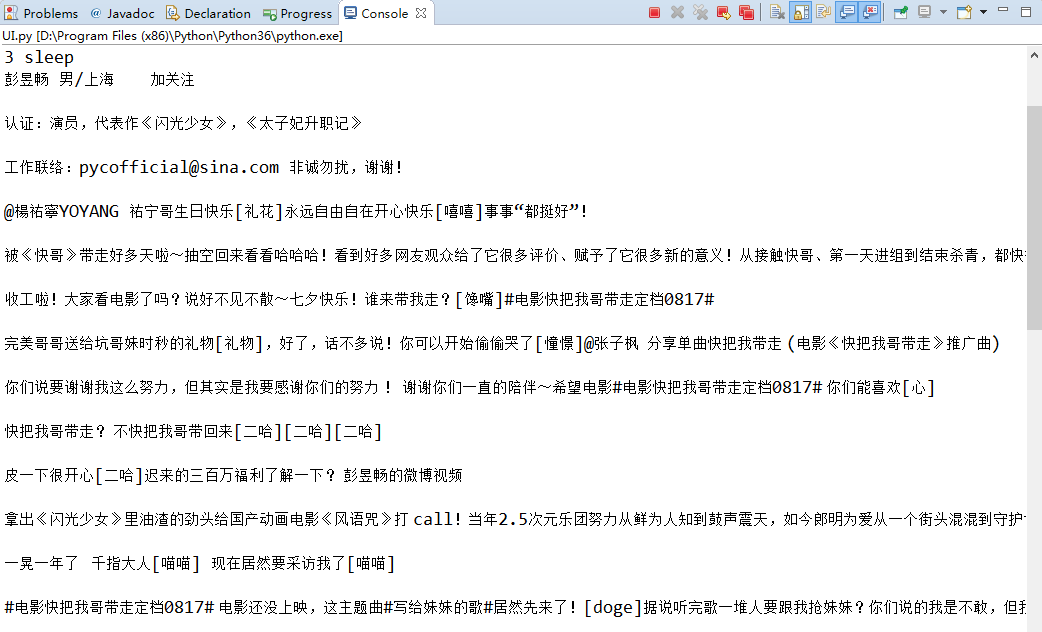
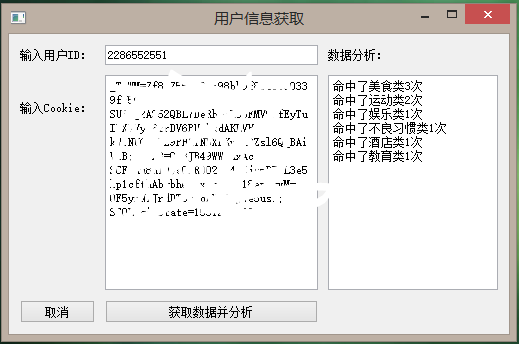


图7 控制台显示爬取数据

最后数据分析函数(analysis(text))利用关键字字典(keywords\_new)对爬取的数据进行分析，最终显示在界面上，如图8所示（对不同用户进行分析）。



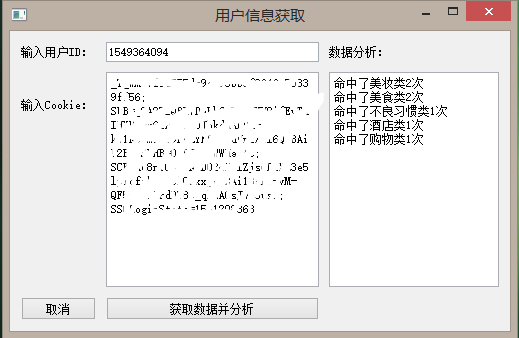


图8 显示结果

## 2.3 容错性

对于错误的输入，程序不运行，显示结果为空，如图9所示。

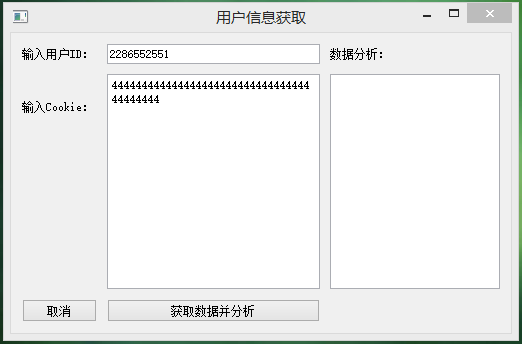


图9 不正确输入

# 3 总结

程序设计的初始想法是结果用饼状图显示，由于对Python和PyQt5的使用属于新手入门，程序运行的一些错误还有待改正，所以结果最终用文本显示。本文的程序对原有的爬虫程序做了如下改进：

（1）使用Python3重新编程，因为Python2和Python3兼容性极差，所以代码改动很大；

（2）添加可视化界面，可轻松爬取数据并查看结果。

还有不足之处有待改进：

（1）最终结果可用图表展示，更加美观和直观；

（2）数据分析种类单一，可增加分析结果；

（3）由于用的是Python3.6编程，难以将项目打包成EXE等可直接运行的文件，还需要借助Python编译器才能执行。

# 参考文献：

[1] 没有人给他写信的程序员. python爬虫爬取指定用户微博图片及内容，并进行微博分类及使用习惯分析，生成可视化图表[OL]. 博客园, 2016. <https://www.cnblogs.com/dmyu/p/6034634.html>

[2] Archi. PyQt5中的布局管理[OL]. 博客园, 2016. https://www.cnblogs.com/archisama/p/5453260.html