****

**计算机学院研究生课程论文**

**（ 2019 至 2020 学年度 第1学期）**

基于Tkinter的图片数据集生成程序

学 院： 计算机科学技术学院

课 程 名 称： 数据库系统实现技术

年 级： 2019级

专 业 名 称： 计算机技术

研 究 方 向： 自然语言处理

学 生 姓 名： 闫世祥

授 课 教 师： 栾静

**新疆师范大学硕士研究生课程论文评阅表**

|  |  |  |  |
| --- | --- | --- | --- |
| **姓名** | **闫世祥** | **专业** | **计算机技术** |
| **年级** | **2019级** | **研究方向** | **自然语言处理** |
| **课程类型** |  | **课程名称** | **数据库系统实现技术** |
| **论文题目** | **基于Tkinter的图片数据集生成程序** | | |
| **授课教师** | **栾静** | **成绩** |  |
| **对论文的评阅意见（不少于150字）**  评阅人签名  年 月 日 | | | |

**摘要**

本课题的主要目的是设计网络面向图片的网络爬虫程序，同时需要满足的是具有一定的性能，考虑到网络爬虫的各种需求。具有一些人性化的可视化操作，即使不懂计算机的人群也能轻松上手，高度的对使用人员透明化。

网络爬虫应用广度优先搜索技术，对于URL进行分析，去重。网络爬虫使用多进程技术，让爬虫具有更强大的抓取能力，对网络爬虫的连接网络设置及读取时间避免无限制的等待。为了适应不同需求，使网络爬虫可以根据预先设定的主题进行爬取。研究网络爬虫原理并实现爬虫的相关功能。

**关键字：**网络爬虫；并发；MySQL数据库；MVC架构

**ABSTRACT**

The main purpose of this project is to design a picture oriented web crawler program, and at the same time, it needs to meet certain performance, considering all kinds of needs of web crawler. With some human-oriented visual operation, even people who do not understand the computer can easily get started, highly transparent to users.

Web crawlers use breadth first search technology to analyze and de duplicate URLs. Web crawlers use multi process technology, so that the crawler has a more powerful grasp ability, and avoid unlimited waiting for the connection network settings and read time of web crawlers. In order to adapt to different needs, the web crawler can crawl according to the preset theme. Study the principle of web crawler and realize the related functions of crawler.

**Keywords**: web crawler; concurrency; MySQL database; MVC architecture

目录

[1 需求分析 1](#_Toc29295729)

[2 可行性分析： 1](#_Toc29295730)

[2.1 技术可行性： 1](#_Toc29295731)

[2.1.1 Python介绍 1](#_Toc29295732)

[2.1.2 MYSQL 介绍 2](#_Toc29295733)

[2.1.3 Tkinter 介绍 3](#_Toc29295734)

[2.1 经济的可行性： 3](#_Toc29295735)

[2.3 社会可行性： 3](#_Toc29295736)

[2.4 风险因素控制的可行性： 3](#_Toc29295737)

[3 详细设计： 4](#_Toc29295738)

[3.1 用例图： 4](#_Toc29295739)

[3.2 健壮性图： 7](#_Toc29295740)

[3.3 顺序图： 8](#_Toc29295741)

[3.4 类图： 8](#_Toc29295742)

[4 实现 9](#_Toc29295743)

[5 测试 9](#_Toc29295744)

[6 总结 12](#_Toc29295745)

[参考文献 13](#_Toc29295746)

# 1 需求分析

随着计算机视觉的迅速发展，各型各色的人工智能产品出现在生活当中。在对于图片的机器学习中用于训练的数据集通常不容易得到，即使有数据通常避免不了手动标签的困扰。试想一下如果有一个网络爬虫系统，可以自动去网上收集需要的数据集并，自动把收集到的图片，文本之类的数据进行自动标签，如此一来避免了很多不必要的麻烦。

图片爬虫系统应运而生，根据输入的关键字，图片数量自动去爬虫生成数据集，保存到本地，文件名作为图片的标签。对于数据库中的图片信息的统计不便，在本系统中需要加入统计模块，以可视化的形式给予用户一个友善的界面。对于有单张图片需求的操作，该系统应该可以合理的处理，对此本系统中添加保存单张图片到本地的功能，可以从远程数据库中查询得到单张图片的URL，查看并保存到本地。

# 2 可行性分析：

## 2.1 技术可行性：

网络爬虫是一个自动提取网页的程序，它为搜索引擎从Web上下载网页，是搜索引擎的重要组成部分。通过网络爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL列表；在抓取网页的过程中，不断从当前页面上抽取新的URL放入待爬行队列，直到满足系统的停止条件。

该系统采用python语言进行开发，利用其丰富的第三方网络爬虫库，可以使得爬虫部分得到实现。对于界面使用python官方的TKINTER，进行构建简单易操作的可视化。数据库采用MySQL，合理利用MySQL的存储语句使该软件的性能可以得到很好的提升。

本文所采用的技术方案是目前基于C/S开源框架--Tkinter，涉及到应用程序开发时使用 Python 语言，数据库使用开源关系型数据库 MySQL，连接数据库技术采用的是Python的开源模块Pymysql。以上采用的技术都是开源基金支持的项目，各自拥有自己成熟、活跃的开源开发者社区，十分有利于本文作者收集相关资料和进行开发技术上的储备。以下主要介绍本系统开发中所用到的主要技术。

### 2.1.1 Python介绍

Python 是一种面向对象、解释型脚本语言，由 Guido van Rossum 于 1989 年发明，1991 年发行第一个公开发行版，现在最新的版本是 Python3.5，并且一直在被维护和更新。Python 语法简洁清晰，它是使用空白符(white space)作为语句缩进，并不需要使用分号。

Python 提供了非常完善的基础代码库，内容覆盖了网络、文件、GUI、数据库、文本等。使用 Python 进行程序开发时，许多功能不必从零编写，直接使用现成的即可。由于 Python 经过这么多年的发展，许多第三方公司的产品为 python提供了接口并且封装成 Python 扩展库了，供开发者直接使用。

Python 是一种完全面向对象的开发语言。Python 常见的数据类型例如函数、模块、数字、字符串都是操作对象，类似与 java 那样支持对象的继承、重载、派生、多继承，极大提高的了代码的扩展性和代码的复用性，同时 Python 还支持重载运算符和动态类型。

Python 语言的设计思想就是一切简单，程序员使用最简单、最少的代码开发用户需要的功能。同样的任何硬币都有两面，任何编程语言都有自己的优势和不足。Python 也有自己的缺点。

第一个缺点就是运行速度相对较慢，这是解释型编程语言的通病。C语言程序在运行前直接编译成计算机能执行的机器语言，所以运行起来速度会非常快。而Python是一种脚本语言，它在执行过程中会被逐行的编译成机器语言，这个编译的过程相对来说是非常耗时的，所以python代码运行速度相对较慢。

第二个缺点就是python代码不能够加密。如果要发布自己的Python程序，实际上就是发布源代码。这一点上C语言会做的很好，C语言不用发布源代码，只需要把编译后的机器语言（Windows上常见的xxx.exe文件）发布出去。而要从机器语言反编译出C代码是几乎不可能实现的。所以，如果是编译型的语言都相对来说安全一些，而解释型的语言则不能够加密，必须要把源码发布出去。

### 2.1.2 MYSQL 介绍

MySQL 数据库简称 MySQL 是一款开源的关系型数据管理系统。最开始是由Apache 基金支持，目前被 Oracle 公司收购。由于 MySQL 数据库体积小、速度快、开源社区成熟、总体拥有成本低等优势 MySQL 受到开发者们的广泛追捧。

MySQL 数据库是基于 Linux 系统开发的开源免费的数据库管理系统，支持 5000万条数据记录，兼容所有的平台，基本上满足了中小企业在数据库上的全部需求，所以 MySQL 数据库成为了时下中小型网站首选数据库。MySQL 所使用的 SQL语言是用于访问数据库的最常用标准化语言。

MySQL 数据库的主要特点：

1. 使用 C 和 C++编写，并使用了多种编译器进行测试，保证了源代码的可移植性。
2. 支持 AIX、FreeBSD、HP-UX、Linux、Mac OS、NovellNetware、OpenBSD、OS/2 Wrap、Solaris、Windows 等多种操作系统。
3. 支持多线程，并行读写数据，提高 CPU 的利用率。
4. 对 SQL 查询算法进行了完善，一定程度上提高了查询的速度。
5. MySQL 可以单独作为一个应用程序对外提供网络数据存储、查询、更新服务，也可以和其他开发程序一起对外提供服务。
6. MySQL 支持多国语言，常见的编码如中文的 GB 2312、BIG5、UTF-8，都可以作为数据表的表名和字段属性名。
7. 支持例如 TCP/IP、ODBC 和 JDBC 等多种协议连接数据库。
8. 拥有功能完善的第三方数据库管理软件例如：Navigate for MySQL
9. 支持大型的数据库，可以处理拥有上千万条记录的大型数据库。
10. 支持多种存储引擎。
11. MySQL 是开源的，所以你不需要支付额外的费用。
12. MySQL 使用标准的 SQL 数据语言形式。

### 2.1.3 Tkinter 介绍

Python的GUI设计工具有Tkinter、Wx Python、PyQt等，各有优缺点。Tkinter是Python自带的标准GUI库，安装Python后直接使用Tkinter，不需额外安装，使用前用from tkinter import\* 导入tkinter。第三方GUI界面设计后经过转换成Python语言才可使用[6]。

Tkinter 是一个 Tk GUI API的标准面向对象接口，最初由Steen Lumholdt 编写，当时 Steen Lumholdt 在忙于利用 Python 改进 GUI。Tkinter 是一个成熟的跨平台接口，它为 GUI 应用程序提供小型的配件集。但是，这并不意味着要固定于这个集合。Tkinter 是可扩展的，即意味着可以使用第三方的配件程序包。Widget是用户接口元素，例如 radio button（单选框）和 list box（列表框）等。

Tkinter 工具箱是一个强大的 GUI 框架，允许 Python 程序运行于 Windows、Unix 和 Macintosh 平台上。Tkinter 和其他工具箱之间的主要不同在于可移植性方面。几乎所有的工具箱都适用于一些特定系统，例如 KDE 绑定 Linux、Mac工具箱绑定 Mac 都是支队一个特定平台提供支持的 GUI 实现，然而，Tkinter 却允许读者编写可在许多平台运行而不用任何改动的代码。

Tkinter证明应用程序编程接口设计可与应用程序的事务历程分开创建。选择Tkinter作为自己的GUI 环境时，主要必须考虑在何处防止正确的配件。完成可视化设计后，只需要把配件操作绑定到需要调用的具体函数上即可，至此图形接口准备停当。Tkinter还允许快速地处理按钮和窗口，并定义其属性。设计和创建自己的接口后，可以改变应用程序的函数而不对GUI代码进行任何改变。

## 2.1 经济的可行性：

由于不涉及收费的软件开发工具，以及过多的开发人员，对于经济的要求非常的小。

## 2.3 社会可行性：

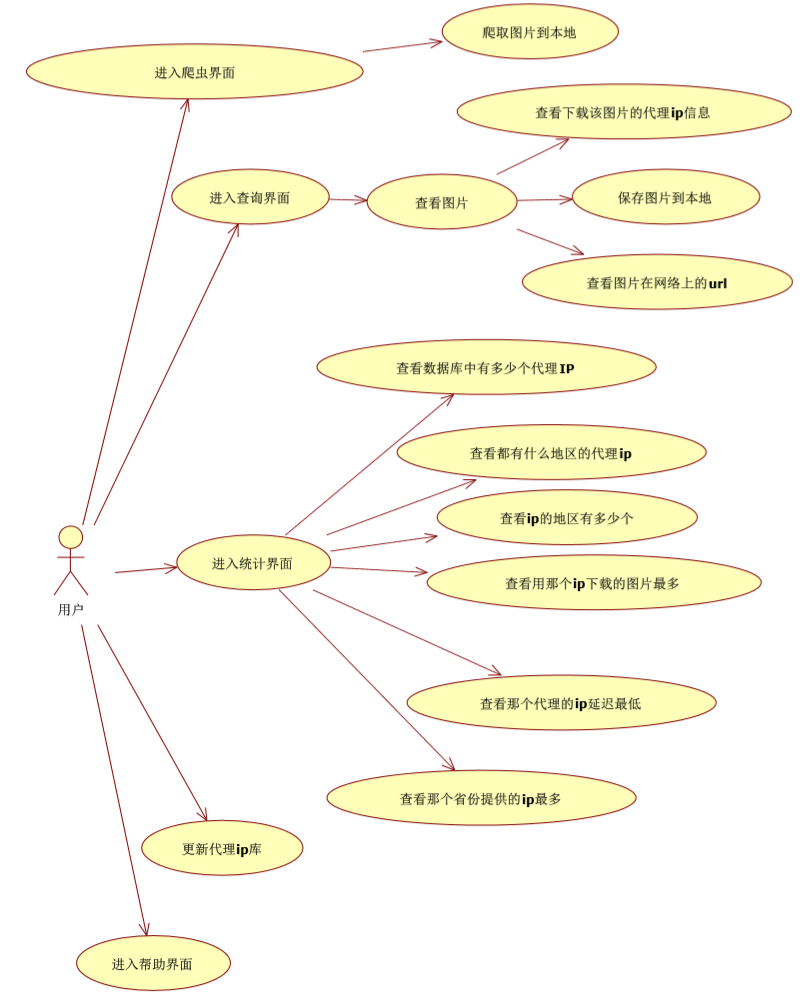
对于爬虫程序国家有严苛的要求，但该系统属于小型爬虫系统，对图片网站的服务器的损害可以忽略不记。该系统有严格的爬虫网站的限制，对于隐私权限的图片、污秽图片、对社会有影响的图片等，不会在该网站中出现。开源软件的使用使得本系统与软件的使用权没有任何利益冲突。

## 2.4 风险因素控制的可行性：

为了限制类似DDOS类的爬虫服务（可以导致服务器瘫痪的访问），防止某些用户设置的图片数额过大并且忘记关闭程序，特意限制爬取一次图片最多不可超过200张，以防止出现不可预料的问题。

# 3 详细设计：

## 3.1 用例图：



用例描述：

表 1 进入爬虫界面

|  |  |
| --- | --- |
| 用例# | 进入爬虫界面 |
| 使用语境 | 用户在主界面鼠标点击爬虫 |
| 范围 | 用户 |
| 触发事件 | 进入爬虫界面 |
| 前置条件 | 进入主页面 |
| 后置条件 | 填写关键字，填写数量，爬取图片 |

表 2 爬取图片到本地

|  |  |
| --- | --- |
| 用例# | 爬取图片到本地 |
| 使用语境 | 用户在爬虫界面点击爬虫 |
| 范围 | 用户 |
| 触发事件 | 将图片爬取到本地，图片信息URL，文件在本地存储的地址存放到数据库中 |
| 前置条件 | 点击爬虫界面 |
| 后置条件 | 执行结束返回爬取结束消息对话框 |

表 3 进入查询界面

|  |  |
| --- | --- |
| 用例# | 进入查询界面 |
| 使用语境 | 用户在主界面鼠标点击查询 |
| 范围 | 用户 |
| 触发事件 | 从数据库中加载查询界面第一个下拉框的选择条目并进入查询界面 |
| 前置条件 | 进入主页面 |
| 后置条件 | 选择图片类别，选择图片，加载图片 |

表 4 查看图片

|  |  |
| --- | --- |
| 用例# | 查看图片 |
| 使用语境 | 在查询界面选择图片类别，选择图片名 |
| 范围 | 用户 |
| 触发事件 | 进入展示图片窗口 |
| 前置条件 | 进入查询页面 |
| 后置条件 | 右击图片进行图片信息的查看以及保存图片到本地 |

表 5 查看图片爬取信息

|  |  |
| --- | --- |
| 用例# | 查看图片爬取信息 |
| 使用语境 | 在展示图片界面右击选择IP信息 |
| 范围 | 用户 |
| 触发事件 | 展示图片的IP信息 |
| 前置条件 | 进入展示图片页面 |
| 后置条件 | 返回消息对话框展示图片的爬取信息 |

表 6 进入统计界面

|  |  |
| --- | --- |
| 用例# | 进入统计界面 |
| 使用语境 | 在展示图片界面右击选择保存 |
| 范围 | 用户 |
| 触发事件 | 从数据库中搜素图片的URL，读取URL表现的数据流到本地的graph文件夹 |
| 前置条件 | 进入展示图片页面 |
| 后置条件 | 返回消息对话框保存图片成功 |

表 7 查看图片爬取信息

|  |  |
| --- | --- |
| 用例# | 查看图片爬取信息 |
| 使用语境 | 在展示图片界面右击选择IP信息 |
| 范围 | 用户 |
| 触发事件 | 展示图片的IP信息 |
| 前置条件 | 进入展示图片页面 |
| 后置条件 | 返回消息对话框展示图片的爬取信息 |

表 8 查看数据库中有多少图片

|  |  |
| --- | --- |
| 用例# | 查看数据库中有多少图片 |
| 使用语境 | 在展示图片界面右击选择IP信息 |
| 范围 | 用户 |
| 触发事件 | 展示图片的IP信息 |
| 前置条件 | 进入展示图片页面 |
| 后置条件 | 返回消息对话框展示图片的爬取信息 |

表 9 查看都有什么地区的代理IP

|  |  |
| --- | --- |
| 用例# | 查看都有什么地区的代理IP |
| 使用语境 | 在展示图片界面右击选择IP信息 |
| 范围 | 用户 |
| 触发事件 | 展示图片的IP信息 |
| 前置条件 | 进入展示图片页面 |
| 后置条件 | 返回消息对话框展示图片的爬取信息 |

表 10 查看IP的地区有多少个

|  |  |
| --- | --- |
| 用例# | 查看IP的地区有多少个 |
| 使用语境 | 在展示图片界面右击选择IP信息 |
| 范围 | 用户 |
| 触发事件 | 展示图片的IP信息 |
| 前置条件 | 进入展示图片页面 |
| 后置条件 | 返回消息对话框展示图片的爬取信息 |

表 11 查看用那个IP下载的图片最多

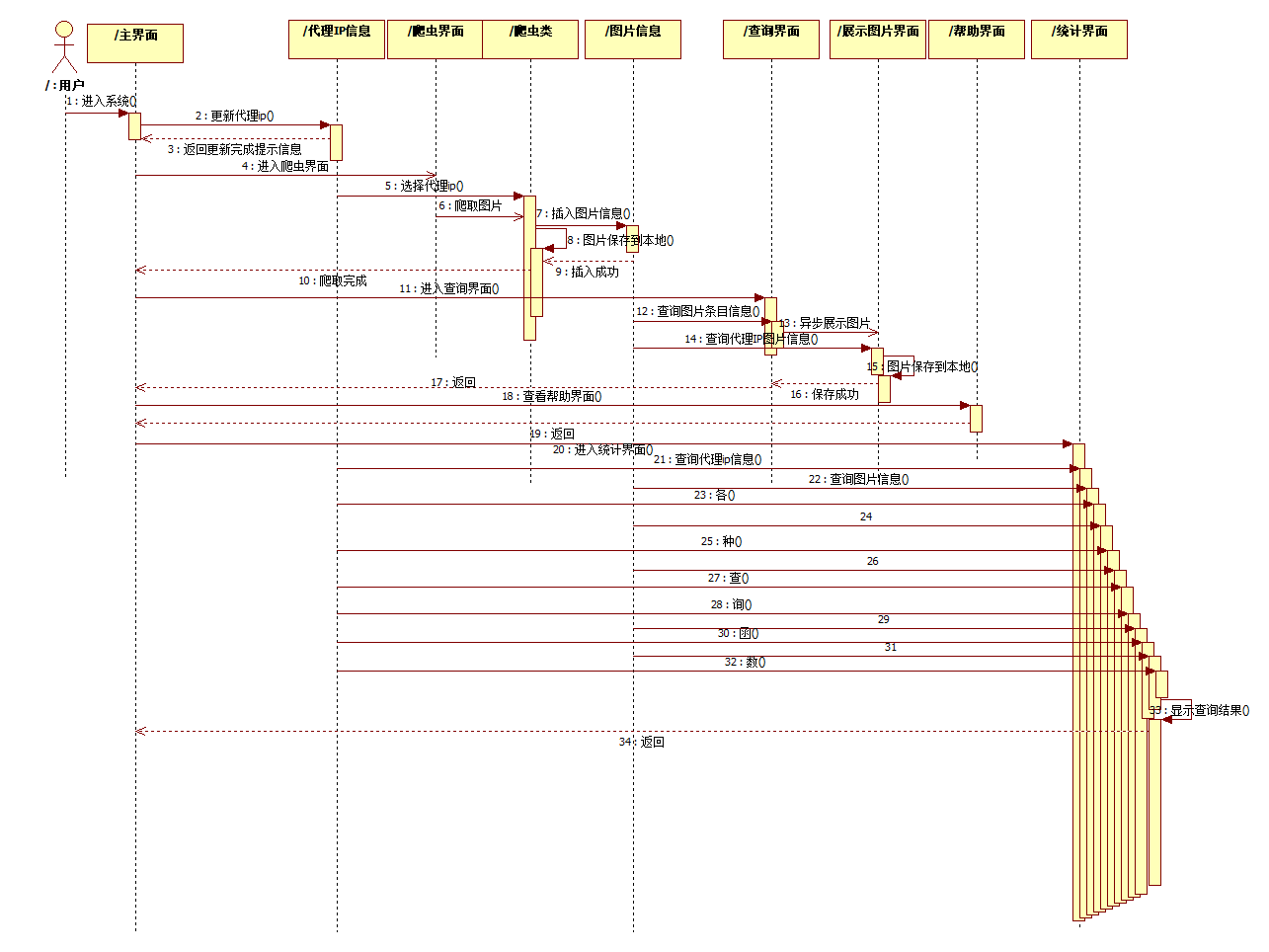
|  |  |
| --- | --- |
| 用例# | 查看用那个IP下载的图片最多 |
| 使用语境 | 在展示图片界面右击选择IP信息 |
| 范围 | 用户 |
| 触发事件 | 展示图片的IP信息 |
| 前置条件 | 进入展示图片页面 |
| 后置条件 | 返回消息对话框展示图片的爬取信息 |

表 12 查看那个代理的IP延迟最低

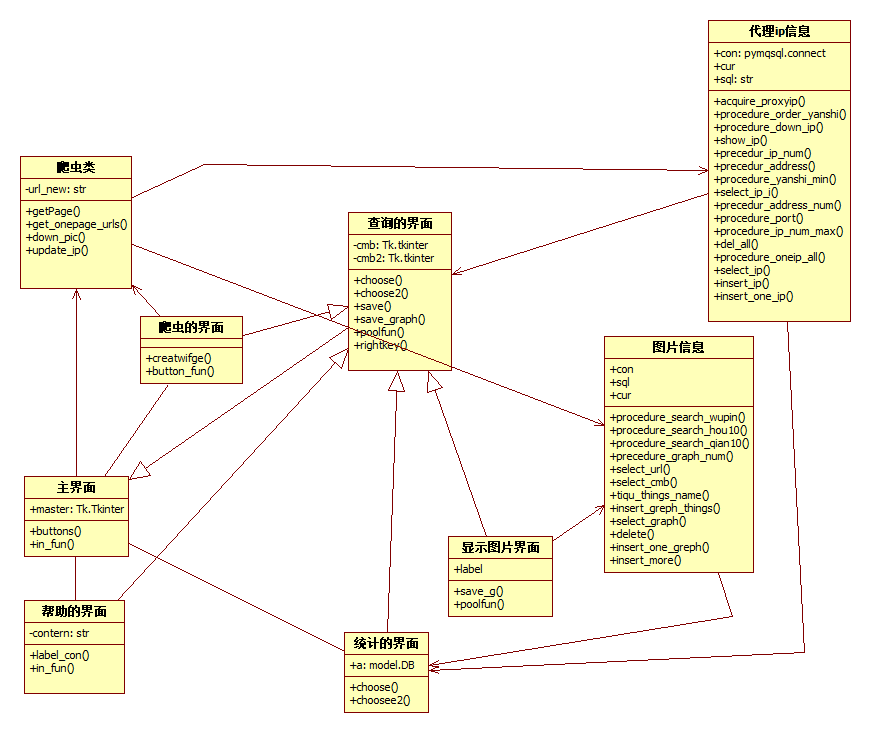
|  |  |
| --- | --- |
| 用例# | 查看那个代理的IP延迟最低 |
| 使用语境 | 在展示图片界面右击选择IP信息 |
| 范围 | 用户 |
| 触发事件 | 展示图片的IP信息 |
| 前置条件 | 进入展示图片页面 |
| 后置条件 | 返回消息对话框展示图片的爬取信息 |

## 3.2 健壮性图：

## 3.3 顺序图：



## 3.4 类图：



# 4 实现

该系统在Win10，Ubantu18.04利用python3.7测试均可有效运行。

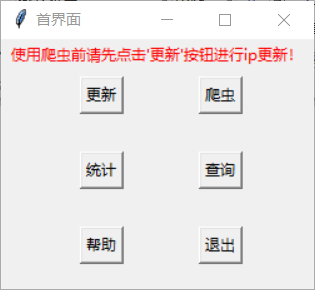
代码实现：详见压缩包，压缩包内包含在window上的可执行程序；以及原代码，在UNIX系统上可自行编译：unzip src.zip&&chmod 777 src&&cd src&& python firstUI.py；系统使用说明书详见首页面帮助。

该系统数据库布置在腾讯云服务器，由于具有公网ipv4地址的便利，使得该软件在任何有网的地方都可以有效使用。

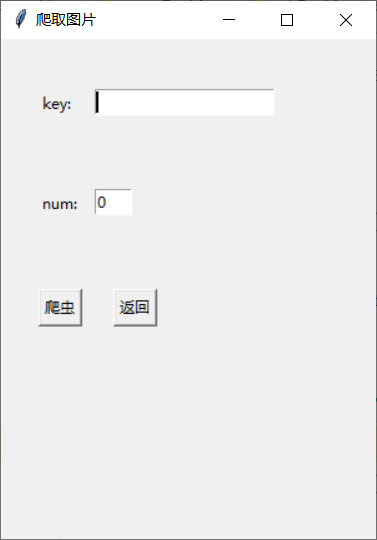
# 5 测试

在python中测试使用语法：“if \_\_name\_\_==’main’”语句。

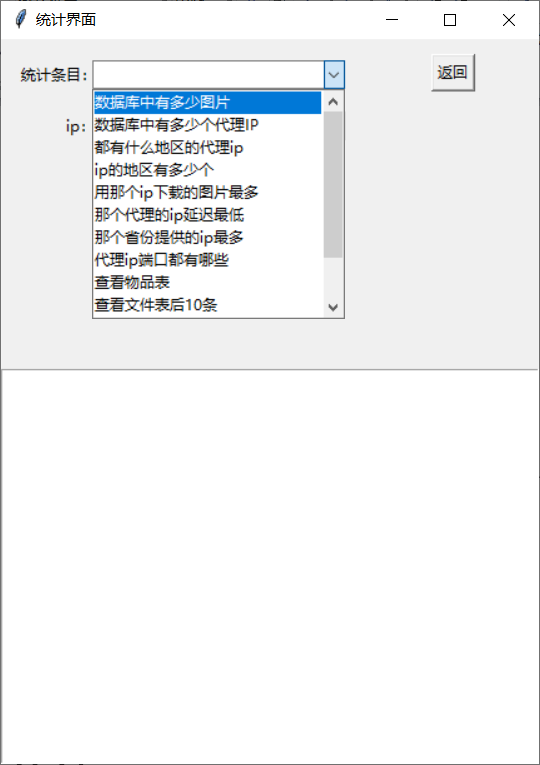
首界面：



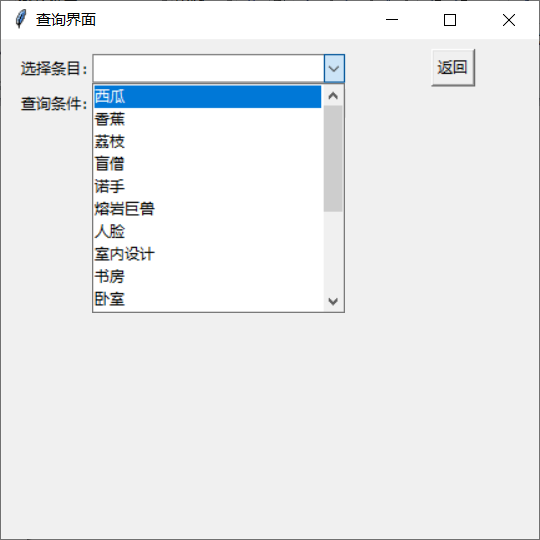
爬虫界面：



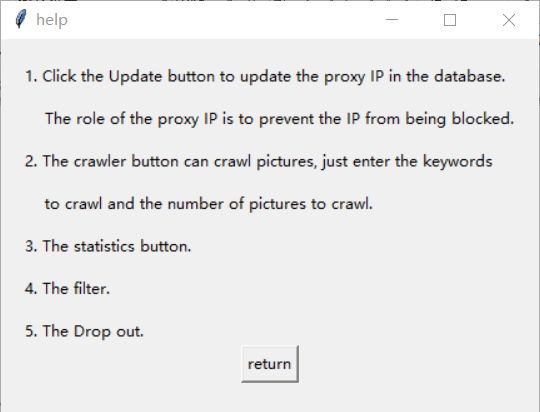
统计界面：



查询界面：



帮助界面：



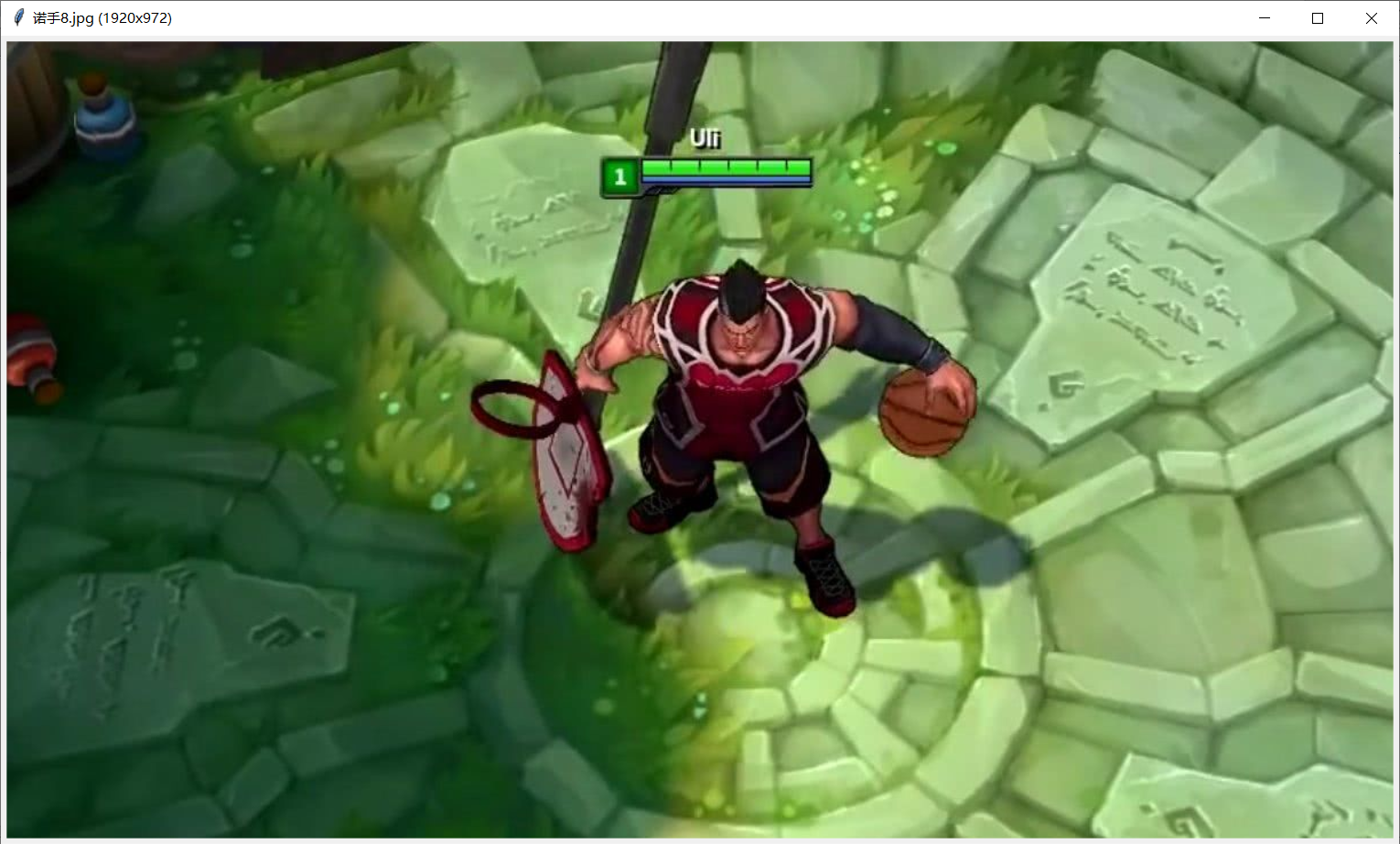
爬虫图片模块：



代理IP模块：



展示图片模块：



# 6 总结

该程序采用python进行编写，基本上实现了生成数据集的功能，但是仍有不足。对于图片保存，仅是保存了网络图片到本地，没有经过任何处理。对于这些直接从网上获得的图片，并没有经过检查，过滤等工作，这样得到的数据集是杂乱无章的，不具有有效性的。应用简单决策树方法，进行减枝操作，对于某些不能启到良好效果的样本数据集进行减枝操作。在以后的工作中，我会添加生成固定尺寸图片的模块，利用到CV2，进行对图片的无损压缩，得到的每张图片分辨率为固定某值的图片。对于某些图片从网络中并不能找到太多样本，利用GAN网络进行生成。

对于并行爬虫，仅做了多进程的处理，并行度不高有待改善。改善方法可用，REDIS数据库做消息调度，用CELERY分布式框架，将该程序部署到多台计算机可大大改善爬虫性

参考文献

[1] Yanqin Ding,Na Li,Jinhan Sun,Linran Zhang,Jianhui Guo,Xueqi Hao,Yuning Sun. Correction to: Oxymatrine Inhibits Bocavirus MVC Replication, Reduces Viral Gene Expression and Decreases Apoptosis Induced by Viral Infection[J]. Virologica Sinica: Official Journal of the Chinese Society for Microbiology,2019,34(6).

[2] Wirtz Nicolas,Dörmann Ulrike,Micke Florian,Filipovic André,Kleinöder Heinz,Donath Lars. Effects of Whole-Body Electromyostimulation on Strength-, Sprint-, and Jump Performance in Moderately Trained Young Adults: A Mini-Meta-Analysis of Five Homogenous RCTs of Our Work Group.[J]. Frontiers in physiology,2019,10.

[3] Ding Yanqin,Li Na,Sun Jinhan,Zhang Linran,Guo Jianhui,Hao Xueqi,Sun Yuning. Correction to: Oxymatrine Inhibits Bocavirus MVC Replication, Reduces Viral Gene Expression and Decreases Apoptosis Induced by Viral Infection.[J]. Virologica Sinica,2019,34(6).

[4] Geiger M,Roche N,Vlachos E,Cattagni T,Zory R. Acute effects of bi-hemispheric transcranial direct current stimulation on the neuromuscular function of patients with chronic stroke: A randomized controlled study.[J]. Clinical biomechanics (Bristol, Avon),2019,70.

[5] Andri Sunardi,Suharjito. MVC Architecture: A Comparative Study Between Laravel Framework and Slim Framework in Freelancer Project Monitoring System Web Based[J]. Procedia Computer Science,2019,157.

[6] M. Geiger,N. Roche,E. Vlachos,T. Cattagni,R. Zory. Acute effects of bi-hemispheric transcranial direct current stimulation on the neuromuscular function of patients with chronic stroke: A randomized controlled study[J]. Clinical Biomechanics,2019,70.

[7] 张哲,吴姗姗.基于J2EE的家政共享系统应用研究[J].现代信息科技,2019,3(22):148-149.

[8] 刘雅慧.基于Struts框架的考研资讯平台的设计与开发[J].现代信息科技,2019,3(22):22-24.

[9] 黄丹雄.校园信息聚合平台的设计与开发[J].现代信息科技,2019,3(21):10-12.

[10] 莫运襟.小议基于SSM的水电站设备管理系统的设计及实现[J].轻工科技,2019,35(12):92-93.

[11] 殷存举.基于ASP.NET MVC技术的在线RSS阅读器的设计与实现[J].信息技术与信息化,2019(11):93-95.

[12] 毋世晓.基于MVC的BBS论坛创建与实现[J].电脑编程技巧与维护,2019(12):21-23.

[13] 孙红艳.基于SSH框架的班级事物管理系统设计与实现[J].电脑编程技巧与维护,2019(12):135-136+139.

[14] 丁浩.基于MVC模式的购物网站设计研究与实现[J].电脑知识与技术,2019,15(33):27-29.

[15] 曹鹏.Spring MVC框架下电信用户信息分析大数据可视化研究[J].电脑知识与技术,2019,15(33):253-254+269.

[16] 吕阳.集装箱综合管理系统体系结构分析与研究[J].电子测试,2019(22):56-57+55.

[17] 杨博文,赵志刚,高阳,冷雪敏.基于MVC框架技术下机组机舱温度智能控制系统研发[J].电子世界,2019(22):64-65.

[18]孙炯宁.高校求职招聘网站系统的设计与实现[J/OL].数字技术与应用:1-3[2020-01-07].http://kns.cnki.net/kcms/detail/12.1369.TN.20191127.1328.008.html.