学校代码：12331

****

毕 业 设 计 说 明 书

基于网络爬虫的机器人助手的设计与实现

|  |  |
| --- | --- |
| **姓 名：** | 高毅 |
| **学 号：** | 170507140113 |
| **指导教：** | 赵学臣（职称） |
| **学 院：** | 数据科学与计算机学院 |
| **专 业：** | 计算机科学与技术 |
| **完成日：** | 2021年4月13日 |

山东女子学院毕业设计作者声明

本人声明：本人完成的毕业设计及呈交的毕业设计说明书是本人在导师指导下取得的成果。对前人及其他人员对毕业设计的启发和贡献已在毕业设计说明书中做出了明确的声明，并表示了谢意。除了特别加以标注和致谢的地方外，本人的毕业设计方案中不包含其他人和其他机构已经实现、发表或者撰写过的成果。

本人同意学校根据《中华人民共和国学位条例暂行实施办法》等有关规定保留本人的毕业设计说明书并向国家有关部门或资料库送交本说明书或其电子版，允许本说明书被查阅或借阅；本人授权山东女子学院可以将本毕业设计说明书的全部或者部分内容编入有关数据库进行检索，可以采用影印、缩印或者其它复制手段汇编本说明书。

作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

摘要

本设计围绕教师成果存储现状，利用爬虫、数据可视化、自动回复机器人，分析了当前需求，解决了教师、学生对教师文献及项目查看不方便的问题，实现了即时查询即时修改的功能，达到了节约时间、方便查看的目的。

本设计的创新之处在于爬虫与C#相结合、利用微信端口设计一个自动回复机器人，可以根据用户发送的关键字来选择回复。为此，采用了Python和C#交互代码、wechat\_pc\_api 使用HOOK技术将核心功能封装成dll，并提供简易的接口给程序调用的方法，而且，运行结果表明，本设计的观点途径是可行的，达到了预期设计目标。

关键词：爬虫；数据可视化；微信扩展端口；机器人

Abstract

This design revolves around the status quo of teacher achievement storage, uses crawlers, data visualization, and automatic reply robots to analyze current needs, solves the inconvenience of teachers and students in viewing teacher documents and projects, and realizes the function of real-time query and real-time modification.

The purpose of saving time and facilitating viewing.

The innovation of this design lies in the combination of crawler and C#, using the WeChat port to design an automatic reply robot, which can choose to reply according to the keywords sent by the user.

To this end, the use of Python and C# interactive code, wechat\_pc\_api use HOOK technology to encapsulate the core functions into dll, and provide a simple interface for the program to call the method, and the running results show that the viewpoint approach of this design is feasible and achieved

Expected design goals.

Key words: spider; data visualization; WeChat expansion port; robot

目录

1 绪论1

1．1关键技术介绍1

1.1.1 Python1

1.1.2 Python爬虫3

1.1.3 C#6

1.1.4 VUE7

1.1.5 sql语句8

2 需求分析4

2.1 课题背景5

2.2 数据获取6

2.3 数据存储6

2.4 用户管理6

2.5 数据展示6

2.6 快捷查询6

3 概要设计4

3.1 系统架构5

3．2 系统环境6

3.3 数据库设计6

4 详细设计4

4.1 Python爬虫设计5

4.2 C#ADO.NET设计6

4.4 Vue.js设计4

4．5 WeChat机器人设计5

5 系统实现4

5.1 5

5.26

5.36

5.46

6 系统测试4

6.15

6

1 绪论

* 1. **关键技术介绍**

**1.1.1 Python**

（1）Python 发展历史

Python 是由 Guido van Rossum 在八十年代末和九十年代初，在荷兰国家数学和计算机科学研究所设计出来的。

现在，全世界差不多有600多种编程语言，但流行的编程语言也就那么20来种。如果你听说过TIOBE排行榜，你就能知道编程语言的大致流行程度。

总的来说，这几种编程语言各有千秋。C语言是可以用来编写操作系统的贴近硬件的语言，所以，C语言适合开发那些追求运行速度、充分发挥硬件性能的程序。而Python是用来编写应用程序的高级编程语言。

（2）Python 特点

1.易于学习：Python有相对较少的关键字，结构简单，和一个明确定义的语法，学习起来更加简单。

2.易于阅读：Python代码定义的更清晰。

3.易于维护：Python的成功在于它的源代码是相当容易维护的。

4.一个广泛的标准库：Python的最大的优势之一是丰富的库，跨平台的，在UNIX，Windows和Macintosh兼容很好。除了内置的库外，Python还有大量的第三方库，供你直接使用的东西。

5.互动模式：互动模式的支持，可以从终端输入执行代码并获得结果的语言，互动的测试和调试代码片断。

6.可移植：基于其开放源代码的特性，Python已经被移植（也就是使其工作）到许多平台。

7.可扩展：如果你需要一段运行很快的关键代码，或者是想要编写一些不愿开放的算法，你可以使用C或C++完成那部分程序，然后从你的Python程序中调用。

(3)适合Python开发的应用

1.首选是网络应用，包括网站、后台服务等；

2.其次是许多日常需要的小工具，包括系统管理员需要的脚本任务；

3.另外就是把其他语言开发的程序再包装起来，方便使用。

(4)Python的缺点

第一个缺点就是运行速度慢，和C程序相比非常慢，因为Python是解释型语言，你的代码在执行时会一行一行地翻译成CPU能理解的机器码，这个翻译过程非常耗时，所以很慢。但是大量的应用程序不需要这么快的运行速度，因为用户根本感觉不出来。

第二个缺点就是代码不能加密。如果要发布你的Python程序，实际上就是发布源代码。凡是编译型的语言，都没有这个问题，而解释型的语言，则必须把源码发布出去。

**1.1.2 基于Python爬虫**

要了解爬虫首先需要了解什么是网络，互联网建立的目的是什么，什么是上网。

(1)什么是互联网

互联网是由网络设备（网线，路由器，交换机，防火墙等等）和一台台计算机连接而成，像一张网一样

(2)互联网建立的目的

互联网的核心价值在于数据的共享/传递：数据是存放于一台台计算机上的，而将计算机互联到一起的目的就是为了能够方便彼此之间的数据共享/传递，否则你只能拿U盘去别人的计算机上拷贝数据了。

(3)爬虫的定义

向网站发起请求，获取资源后分析并提取有用数据的程序。

(4)爬虫的分类

1.通用爬虫

2.聚焦爬虫

3.增量式爬虫

本设计目标网站需要登录即不想让爬虫爬取，因此需要使用聚焦爬虫。

(5)爬虫的基本流程

保存数据

解析内容

获取响应内容

发送请求

1、发送请求

使用http库向目标站点发起请求，即发送一个Request。

Request包含：请求头、请求体等。

2、获取响应内容

如果服务器能正常响应，则会得到一个Response

Response包含：html，json，图片，视频等。

3、解析内容

解析html数据：正则表达式，第三方解析库如Beautifulsoup，pyquery等。

解析json数据：json模块

解析二进制数据:以b的方式写入文件

1. 保存数据：数据库、文件

(6) 爬虫requests的组成

1、请求方式：

常用的请求方式：GET，POST

post请求的参数放在请求体内，可用浏览器查看，存放于form data内，get请求的参数直接放在url后。

2、请求url

url全称统一资源定位符，如一个网页文档，一张图片，一个视频等都可以用url唯一来确定。

url编码 [https://\*\*\*/\*?wd=\*\*](https://***/*?wd=**)

3、 请求头

User-agent：请求头中如果没有user-agent客户端配置，

服务端可能将你当做一个非法用户

host

cookies：cookie用来保存登录信息，存放在本地

4、data

用来存放模拟登录时所需的用户数据

User、Password等一系列数据。

(7) Response 的组成

1、响应状态

200代表成功，301代表跳转，404文件不存在，403权限，502服务器错误。

2、Response header

set-cookie：可能有多个，是来告诉浏览器，把cookie保存下来

3、preview就是网页源代码最主要的部分，包含了请求资源的内容，如网页html、图片、二进制数据等。

**1.1.3 C#**

（1）C#发展史

C#是微软公司在2000年6月发布的一种新的编程语言, C#是面向对象的语言，而且它是第一个面向组件的编程语言。

（2）C#特点

C#是由C和C++衍生出来的面向对象的编程语言，运行于.NET Framework和.NET Core(完全开源，跨平台)之上的高级程序设计语言。C#看起来与Java有着惊人的相似；它包括了诸如单一继承、接口、与Java几乎同样的语法和编译成中间代码再运行的过程。但是C#与Java有着明显的不同，它借鉴了Delphi的一个特点，与COM（组件对象模型）是直接集成的，而且它是微软公司 .NET windows网络框架的主角。它在继承C和C++强大功能的同时去掉了一些它们的复杂特性。

（3）ADO.NET

ADO.NET是一个COM组件库，用于访问数据库的接口。

（4）ADO.NET类库

1. Connection 类

和数据库交互，必须连接它。连接帮助指明数据库服务器、数据库名字、用户名、密码，和连接数据库所需要的其它参数。Connection对象会被Command对象使用，这样就能够知道是在哪个数据源上面执行命令。与数据库交互的过程意味着必须指明想要执行的操作。这是依靠Command对象执行的。开发人员使用Command对象来发送SQL语句给数据库。Command对象使用Connection对象来指出与哪个数据源进行连接。开发人员能够单独使用Command对象来直接执行命令，或者将一个Command对象的引用传递给DataAdapter，它保存了一组能够操作下面描述的一组数据的命令。

2. Command对象

成功与数据建立连接后,就可以用Command对象来执行查询、修改、插入、删除等命令；Command对象常用的方法有ExecuteReader（）方法、ExecuteScalar（）方法和ExecuteNonQuery（）方法；插入数据可用ExecuteNonQuery（）方法来执行插入命令。

3. DataReader类

许多数据操作要求开发人员只是读取一串数据。DataReader对象允许开发人员获得从Command对象的SELECT语句得到的结果。考虑性能的因素，从DataReader返回的数据都是快速的且只是“向前”的数据流。这意味着开发人员只能按照一定的顺序从数据流中取出数据。这对于速度来说是有好处的，但是如果开发人员需要操作数据，更好的办法是使用Dataset。

4. DataSet对象

DataSet对象是数据在内存中的表示形式。它包括多个DataTable对象，而DataTable包含列和行，就像一个普通的数据库中的表。开发人员甚至能够定义表之间的关系来创建主从关系（parent-child relationships）。DataSet是在特定的场景下使用――帮助管理内存中的数据并支持对数据的断开操作的。DataSet是被所有Data Providers使用的对象，因此它并不像Data Provider一样需要特别的前缀。

5. DataAdapter类

某些时候开发人员使用的数据主要是只读的，并且开发人员很少需要将其改变至底层的数据源。同样一些情况要求在内存中缓存数据，以此来减少并不改变的数据被数据库调用的次数。DataAdapter通过断开模型来帮助开发人员方便的完成对以上情况的处理。当在一单批次的对数据库的读写操作的持续的改变返回至数据库的时候，DataAdapter 填充（fill）DataSet对象。DataAadapter包含对连接对象以及当对数据库进行读取或者写入的时候自动的打开或者关闭连接的引用。另外，DataAdapter包含对数据的SELECT、INSERT、UPDATE和DELETE操作的Command对象引用。开发人员将为DataSet中的每一个Table都定义DataAadapter，它将为开发人员照顾所有与数据库的连接。所以开发人员将做的工作是告诉DataAdapter什么时候装载或者写入到数据库。

**1.1.4 VUE.JS**

Vue是一套用于构建用户界面的渐进式框架，与其它大型框架不同的是，Vue被设计为可以自底向上逐层应用。Vue的核心库只关注试图层，不仅易于上手，还便于与第三方库或既有项目整合。另一方面，当与现代化的工具链以及各种支持类库结合使用时，Vue也完全能够为复杂的单页应用提供驱动。

1. 特点

易用、灵活、性能

（2）结构

|  |  |
| --- | --- |
| 目录/文件 |  |
| build | 项目构建(webpack)相关代码 |
| config | 配置目录，包括端口号等。我们初学可以使用默认的。 |
| node\_modules | npm 加载的项目依赖模块 |
| src | 包含了几个目录及文件： |
| static | 静态资源目录，如图片、字体等。 |
| test | 初始测试目录，可删除 |
| .xxxx文件 | 这些是一些配置文件，包括语法配置，git配置等。 |
| index.html | 首页入口文件，你可以添加一些 meta 信息或统计代码啥的。 |
| package.json | 项目配置文件 |
| README.md | 项目的说明文档，markdown 格式 |

**1.1.5 sql语句**

(1)sql简介

SQL 是用于访问和处理数据库的标准的计算机语言，利用Sql语句可以实现对数据库的各项操作。

(2)sql功能

SQL面向数据库执行查询

SQL 可从数据库取回数据

SQL 可在数据库中插入新的记录

SQL 可更新数据库中的数据

SQL 可从数据库删除记录

SQL 可创建新数据库

SQL 可在数据库中创建新表

SQL 可在数据库中创建存储过程

SQL 可在数据库中创建视图

SQL 可以设置表、存储过程和视图的权限

(3)sql语法

SQL DML 和 DDL

可以把 SQL 分为两个部分：数据操作语言 (DML) 和 数据定义语言 (DDL)。

SQL (结构化查询语言)是用于执行查询的语法。但是 SQL 语言也包含用于更新、插入和删除记录的语法。

查询和更新指令构成了 SQL 的 DML 部分：

SELECT - 从数据库表中获取数据

UPDATE - 更新数据库表中的数据

DELETE - 从数据库表中删除数据

INSERT INTO - 向数据库表中插入数据

SQL 的数据定义语言 (DDL) 部分使我们有能力创建或删除表格。我们也可以定义索引（键），规定表之间的链接，以及施加表间的约束。

SQL 中最重要的 DDL 语句:

CREATE DATABASE - 创建新数据库

ALTER DATABASE - 修改数据库

CREATE TABLE - 创建新表

ALTER TABLE - 变更（改变）数据库表

DROP TABLE - 删除表

CREATE INDEX - 创建索引（搜索键）

DROP INDEX - 删除索引

(4)部分重要语句

1.SQL SELECT 语句：

SELECT 语句用于从表中选取数据, 结果被存储在一个结果表中（称为结果集）。

SQL SELECT 语法: SELECT 列名称 FROM 表名称以及SELECT \* FROM 表名称。

2.INSERT INTO 语句：

INSERT INTO 语句用于向表格中插入新的行。

INSERT INTO 语句用法：

INSERT INTO 表名称 VALUES (值1, 值2,....)，我们也可以指定所要插入数据的列：INSERT INTO table\_name (列1, 列2,...) VALUES (值1, 值2,....)

3. Update 语句

Update 语句用于修改表中的数据。

Update 语句 语法：UPDATE 表名称 SET 列名称 = 新值 WHERE 列名称 = 某值。

4. DELETE 语句

DELETE 语句用于删除表中的行。

DELETE 语句 语法：DELETE FROM 表名称 WHERE 列名称 = 值

2 需求分析

**2.1 课题背景**

现如今，计算机行业发展较快，每一位计算机行业的从业者都拥有属于自己的论文、项目等成果，每一位从事计算机教育事业的教师也不例外，每一位教师都拥有大量的成果，因此，如何较为合理的存储，方便快捷的查询就成了必须解决的问题。

**2.2 数据获取**

所有数据来自学校网站中存储的数据。通过提供的网站进行数据获取，登陆时需对验证码进行处理，在登陆时自动输入到data中实现自动登录。登录成功后进入相关页面对目标数据进行爬取。

**2.3 数据存储**

将Python爬取的数据存放到SQL数据库中，方便后续使用sql语句进行查询、修改等操作。

**2.4 用户管理**

每个用户都可以通过前端进行注册，注册后可以进行登录，登录成功后可以查看属于该用户的数据。

**2.5 数据展示**

数据库存储的数据需要在前端利用Vue.js以表格形式进行可视化展示。

**2.6 快捷查询**

用户可以通过微信向We

3 概要设计

**3.1 系统架构**

该系统主要分为三部分：

（1）用Python的爬虫进行数据获取

（2）基于C#中ADO.NET以及sql语句实现对数据库的各项操作、Vue.js实现用户注册、登录、可视化。

（3）WeChat机器人实现对数据快速灵活查询访问。

系统示意图如下：

图示

描述已自动生成

**3.2 系统开发环境**

整个系统开发环境为win10 PC机 所需环境变量有Python path,

VSCODE path, TESSDATA\_PREFIX : py\Scripts\tessdata。

开发者工具有：（1）用于编写Python爬虫的Spyder和Visual studio。

(2)用于编写C# ADO.NET的Visual studio。

(3)用于设计前端Vue.js的Visual Studio Code。

**3.3 数据库设计**

本系统采用的是SQL Sever 数据库，共需三个表，分别为：login(详见表3-1), userdata(详见表3-2), data(详见表3-3)。

以下为表的详细信息：

|  |  |
| --- | --- |
| username | 注册时的用户名 主键，不可为空 |
| password | 注册时输入的密码 不可为空 |

表3-1 login

Create table login (username int not null, password char(8) not null)

|  |  |
| --- | --- |
| username | 用户名 |
| id | 姓名 |
| part | 角色 |
| Pnum | 手机号 |
| Postnum | 邮箱 |
| sex | 性别 |

表3-2 userdata

Create table userdata(username int notnull, id nchar(8) notnull,part int notnull,pnum vchar(11),postnum,vchar(18), sex char(2))

|  |  |
| --- | --- |
| fname | 文件名称 |
| name | 成员名称 |
| type | 文件类型 |
| date | 完成时间 |

表3-3 data

Create table data(fname vchar(30) primary key not null, name vhar(30) not null, type vchar(5) not null ,data datetime not null)

4 详细设计

**4.1 数据获取-爬虫**

针对爬虫这一部分，首先想到的是使用Python语言。基于Python的爬虫，最主要的就是requests库，requests库包含以下几种方法：

requests.get(url)

requests.post(url)

requests.put(url)

requests.delete(url)

requests.head(url)

requests.options(url)

通过查看网站审查元素找到request请求头中的方法就可以查看该网站的请求方法，查看后该网站的请求方法为requests.get()。

由于网站爬取数据前需要登录，因此需要一个模拟登录的过程。登录时所需数据也可在请求头处查看，将所需数据填写到模拟登录数据中即可实现自动登录，每次爬取数据时自动登录就可以成功的将登陆后数据爬取。

在登录时需要输入验证码，因此需要另外解决此问题。对这个问题，我采用的是用tesserocr识别图像验证码。Tesserocr库：是Python的一个OCR库，安装后可通过对下载图片的灰度处理和二级化处理进行机器识别来获取验证码

**5 系统实现**

**6 系统测试**