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…………………………………………………………………………………6

1 关键技术介绍……………………………………………………………… 6

1.1.1 Python……………………………………………………………… 6

1.1.2基于Python的爬虫……………………………………………… 7

1.1.3 C#…………………………………………………………………8

1.1.4 ADO.NET……………………………………………………………8

1 绪论

* 1. **关键技术介绍**

**1.1.1 Python**

①Python 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。

②Python 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。

④Python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

1. 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。

1. 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。

1. 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。

1. 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

4、保存数据：数据库、文件

1. 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。

1. 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#**