

**说明文档**

课程名称 Python语言程序设计

题 目 网络爬虫设计——豆瓣网站

专业名称 信息工程学院

班 级 18计算机应用技术

姓 名 魏玥

学 号 1120180767

**目 录**

[1.文档概述](#_Toc11529_WPSOffice_Level1) [1](#_Toc11529_WPSOffice_Level1)

[1.1.目的](#_Toc21709_WPSOffice_Level2) [1](#_Toc21709_WPSOffice_Level2)

[1.2 定义、首字母缩写词和缩略语](#_Toc2039_WPSOffice_Level2) [1](#_Toc2039_WPSOffice_Level2)

[1.3 运行范围](#_Toc11526_WPSOffice_Level2) [1](#_Toc11526_WPSOffice_Level2)

[1.4 参考资料](#_Toc4806_WPSOffice_Level2) [1](#_Toc4806_WPSOffice_Level2)

[1.5 概述](#_Toc11058_WPSOffice_Level2) [1](#_Toc11058_WPSOffice_Level2)

[2. 整体说明](#_Toc9334_WPSOffice_Level1) [2](#_Toc9334_WPSOffice_Level1)

[2.1 主要技术说明](#_Toc21762_WPSOffice_Level2) [2](#_Toc21762_WPSOffice_Level2)

[3.具体模块说明](#_Toc648_WPSOffice_Level1) [3](#_Toc648_WPSOffice_Level1)

[3.1书籍信息抓取模块](#_Toc16360_WPSOffice_Level2) [3](#_Toc16360_WPSOffice_Level2)

[3.2 电影信息抓取模块](#_Toc27553_WPSOffice_Level2) [4](#_Toc27553_WPSOffice_Level2)

[3.3 音乐信息抓取模块](#_Toc23511_WPSOffice_Level2) [4](#_Toc23511_WPSOffice_Level2)

[4.项目需求中需考虑的问题](#_Toc5192_WPSOffice_Level1) [6](#_Toc5192_WPSOffice_Level1)

[4.1数据库的设计](#_Toc26157_WPSOffice_Level2) [6](#_Toc26157_WPSOffice_Level2)

[4.2故障处理能力](#_Toc31622_WPSOffice_Level2) [6](#_Toc31622_WPSOffice_Level2)

[5．性能需求](#_Toc25676_WPSOffice_Level1) [6](#_Toc25676_WPSOffice_Level1)

[5.1 准确性和及时性](#_Toc4530_WPSOffice_Level2) [6](#_Toc4530_WPSOffice_Level2)

[5.2 开放性和可扩充性](#_Toc27594_WPSOffice_Level2) [6](#_Toc27594_WPSOffice_Level2)

[5.3 先进性](#_Toc10839_WPSOffice_Level2) [7](#_Toc10839_WPSOffice_Level2)

[5.4 响应速度快](#_Toc27939_WPSOffice_Level2) [7](#_Toc27939_WPSOffice_Level2)

1.文档概述

互联网与人们的生活紧密相连。随着移动互联网时代的到来，数据十分重要，所以利用Python编写网络爬虫技术最具潜力，在市场的地位也很重要。如何利用Python实现豆瓣网站信息的获取成为主要问题。

## 1.1.目的

(1) 本需求分析说明书是为了实现网络爬虫技术而编写的，主要面向分析员、程序员、测试员和最终用户，使之能更好地理解系统功能以及实现系统。

(2) 本说明书不仅是整个软件开发的依据，对以后的维护性工作也起着指导性的作用，同时也是测试计划和用户手册的编写依据。

## 1.2 定义、首字母缩写词和缩略语

无

## 1.3 运行范围

（1）考虑处理器型号及内存容量；

（2）考虑外存容量、联机或脱机、媒体及其存储格式，设备的型号及数量；（3）输入及输出设备的型号和数量，联机或脱机；

（4）数据通信设备的型号和数量；

（5）功能键及其他专用硬件；

（6）服务器端在win7 32位

（7）软件限制: 应安装杀毒软件，防止病毒干扰。

## 1.4 参考资料

1. 郭丽蓉. 基于Python的网络爬虫程序设计[J].电子技术与软件程,2017,000(023)：248-249.
2. 王娟，吴金鹏. 网络爬虫的设计与实现[J].软件导刊.2012,011(004):136-137.

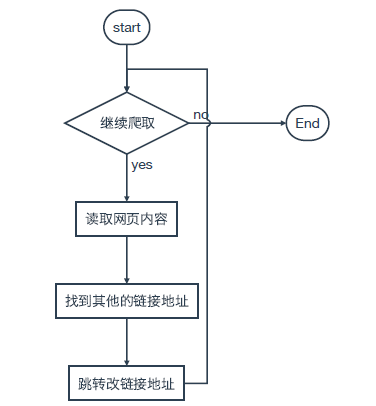
## 1.5 概述

本系统旨在针对需求者以及需要相关信息服务的网民建设网络爬虫设计系统利用python技术进行爬取豆瓣书籍、电影、音乐的信息，并使用数据库对所爬取的数据进行保存与读取，爬取的数据超过1000条以上。

# 2. 整体说明

系统大致划分成四个模块：书籍信息抓取模块、电影信息抓取模块、音乐信息抓取模块、数据保存模块。

抓取数据流程图如下，不断循环，直到将网站上所有的网页吃光。



## 2.1 主要技术说明

网络爬虫意思是：自动抓取万维网信息的程序或脚本。在本系统中利用python技术进行爬取书影音的信息。使用python时，主要使用urllib2（python获取URL的一个组件）以及requests库。在使用requests库过程当中使用到了Beautiful Soup包。它是一个XML的解析器（由Python撰写），并且它可以生成相应的剖析树(parse tree)。如：

bs = Beautiful Soup(文件名.encode('utf-8'),'lxml')

此行代码意思是：将网页信息解析成lxml形式，bs就是Beautiful Soup处理格式化后的字符串。

网络爬虫需要注意的一点是：用不同的浏览器在发送请求时，会有不同的User-Agent头，所以加上User-Agent，伪装成一个被公认的浏览器。除此之外，因为各大网站都有防止爬取的措施，所以在实现网络爬虫时需要有代理IP以及利用sleep包，设置延时时间，防止爬取过于频繁而导致被封IP号。由于本人在一开始缺少该方面的知识，导致IP被封了好几次。

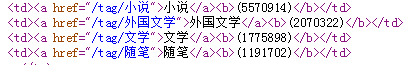
# 3.具体模块说明

## 3.1书籍信息抓取模块

主要分为两个部分：抓取书籍分类链接，然后通过分类链接抓取书籍信息。分别用两个py文件实现。

1、第一步，抓取书籍分类链接，先进入[豆瓣读书](https://www.baidu.com/s?wd=%E8%B1%86%E7%93%A3%E8%AF%BB%E4%B9%A6&tn=24004469_oem_dg&rsv_dl=gh_pl_sl_csd" \t "https://blog.csdn.net/weixin_37656939/article/details/_blank)的分类目录   
 [https://book.douban.com/tag/?view=type&icn=index-sorttags-all](https://book.douban.com/tag/?view=type&icn=index-sorttags-all" \t "https://blog.csdn.net/weixin_37656939/article/details/_blank)

我们需要把这个网站上的分类链接爬取下来，查看页面源代码如下，可以发现对应的链接在a标签下。



可利用两种方式获取a标签中的内容：

* 利用正则表达式，返回pattern实例，根据规则获取匹配字符串，如：

pattern = re.compile(r'<a href="/tag/\W\*">',re.S)

* 利用CSS选择器，其中使用到了使用到了Beautiful Soup包。

抓取相应的超链接如：

<https://book.douban.com/tag/小说>

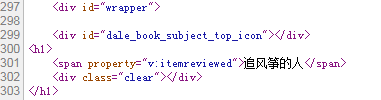
<https://book.douban.com/tag/外国文学>

<https://book.douban.com/tag/文学>

将对应的书籍类型存入数据库（书籍类型表），并将抓取的超链接存入本地文件中。

2、第二步、抓取书籍信息，根据文本文档中的相关链接进入相关页面，如： <https://book.douban.com/tag/文学>

由于类型很多，所以每一个分类里面，只爬取第一页的内容。进入每一本书的详细内容页面，观察页面源代码如下：



根据上面说的CSS选择器的方法，可以得到书名，作者，译者，评价人数，评分，还有这本书的封面链接和简介，如：

title = bookSoup.select('#wrapper > h1 > span')[0].contents[0]

以此类推，将抓取到的内容存入数据库中（书籍信息表），根据（书籍类型表）将书籍类型ID存入（书籍信息表）中。

可利用urllib.urlretrieve(coverUrl,tag);下载封面图片。

## 3.2 电影信息抓取模块

与抓取书籍信息不同，由于电影没有固定的分类。所以本次抓取豆瓣电影榜单top250。

内容共有10页，每页共有25条记录。前三页的网址如下所示：

<https://movie.douban.com/top250?start={}&filter=>

<https://movie.douban.com/top250?start=25&filter=>

<https://movie.douban.com/top250?start=50&filter=>

* 根据对应的规律抓取各自的电影信息，可得到以下的表达式获取每页URL：

url = 'https://movie.douban.com/top250?start={}&filter='.format(i\*25)

* 然后进入每一部电影的详细内容页面，观察页面源代码如下：



根据上面说的CSS选择器的方法，可以得到相关的电影内容，并将内容存入数据库中（电影信息表）。

## 3.3 音乐信息抓取模块

主要分为两个部分：抓取音乐分类链接，然后通过分类链接抓取音乐信息。分别用两个py文件实现。

1、第一步，抓取音乐分类链接，先进入[豆瓣音乐](https://www.baidu.com/s?wd=%E8%B1%86%E7%93%A3%E8%AF%BB%E4%B9%A6&tn=24004469_oem_dg&rsv_dl=gh_pl_sl_csd" \t "https://blog.csdn.net/weixin_37656939/article/details/_blank)的分类目录   
 <https://music.douban.com/tag/>

我们需要把这个网站上的分类链接爬取下来，查看页面源代码如下，可以发现对应的链接在a标签下。



抓取相应的超链接如：

<https://music.douban.com/tag/OST>

<https://music.douban.com/tag/pop>

<https://music.douban.com/tag/民谣>

将对应的音乐类型存入数据库（音乐类型表），并将抓取的超链接存入本地文件中。

2、第二步、抓取音乐信息，根据文本文档中的相关链接进入相关页面，如： <https://music.douban.com/tag/OST>

由于类型很多，所以每一个分类里面，只爬取第一页的内容。进入每一首音乐的详细内容页面，观察页面源代码如下：



根据上面说的CSS选择器的方法，可以得到相关的音乐内容。

以此类推，将抓取到的内容存入数据库中（音乐信息表），根据（音乐类型表）将音乐类型ID存入（音乐信息表）中。

# 4.项目需求中需考虑的问题

## 4.1数据库的设计

（1）服务器程序在管理员与用户进行各种操作（创建，修改，删除，查看，删除用户）时需要对数据库进行数据提交，即对数据库进行查询与修改：在上面的操作过程中都需要对数据库中的所有表，进行联合查询、修改。

（2）接口传递的信息是用数据结构封装了的数据，以参数传递或返回值的形式在各模块间进行传输。

（3）对于出错信息要返回出错提示给用户，并提供出错信息的处理方法以保证软件运行的可维护性。

## 4.2故障处理能力

系统的操作主要是对数据库的管理和维护，其中发生的故障也是针对数据库的，基于数据库有日志文件，可以很方便的修复；由于目前开发受限，服务器可能会出现无响应等行为，因此，对该方面需要做到很好的处理。

# 5．性能需求

## 5.1 准确性和及时性

（1）系统处理的准确性和及时性是系统的必要性能。

（2）在系统设计和开发过程中，要充分考虑系统当前和将来可能承受的工作量，使系统的处理能力和响应时间能够满足用户对信息处理的需求。

（3）由于系统的查询功能对于整个系统的功能和性能完成举足轻重。作为系统的很多数据来源，而模板的种类又影响用户的决策活动，其准确性很大程度上决定了系统的成败。在系统开发过程中，必须采用一定的方法保证系统的准确性。

## 5.2 开放性和可扩充性

## （1）系统在开发过程中，应该充分考虑以后的可扩充性。用户的需求也会不断的更新和完善。所有这些，都要求系统提供足够的手段进行功能的调整和扩充。

（2）要实现可扩充性，应通过系统的开放性来完成，即系统应是一个开放系统，只要符合一定的规范，可以简单的加入或减少系统的模块。

## 5.3 先进性

（1） 通过系统的开放性和可扩充性，不断改善系统的功能完成。

（2） 在系统设计和开发的过程中，应在考虑成本的基础上尽量采用当前主流并有良好发展前途的产品。

## 5.4 响应速度快

（1） 系统在日常处理中的响应速度为秒级，达到实时要求以及时反馈信息。

（2）进行统计分析时，根据所需数据量的不同而从秒级到分钟级，则是保证工作人员不会因为速度问题而影响工作效率。