**中北大学**

**python 大数据实验室**

**题目:**

**利用BS4库爬取豆瓣TOP250网页相关信息**

**姓名:王雁强**

**目录**

[一.项目介绍 3](#_Toc3588615)

[二.依赖库介绍 3](#_Toc3588616)

[2.1 Urllib3 3](#_Toc3588617)

[2.2 Random 3](#_Toc3588618)

[2.3 Time 3](#_Toc3588619)

[2.4 Beautifulsoup4 3](#_Toc3588620)

[2.5 Xlwt 4](#_Toc3588621)

[三.环境配置 4](#_Toc3588622)

[3.1操作系统 4](#_Toc3588623)

[3.2 python版本 4](#_Toc3588624)

[3.3编辑器 4](#_Toc3588625)

[3.4 库的安装 4](#_Toc3588626)

[四. 执行网络爬虫前的准备 5](#_Toc3588627)

[4.1目标网页介绍 5](#_Toc3588628)

[4.2 爬虫前的其他代码详解 6](#_Toc3588629)

[五. 爬虫本体部分的代码讲解 8](#_Toc3588630)

[5.1 爬取免费代理网站并获取完整IP地址 8](#_Toc3588631)

[5.2 使用 beautifulsoup4库 对豆瓣网站TOP250电影相关信息进行爬取 9](#_Toc3588632)

[六.爬取信息之后数据存储 10](#_Toc3588633)

[6.1 将数据存储进入Excel文件中 10](#_Toc3588634)

[6.2 将文件信息存入数据库 11](#_Toc3588635)

[6.2.1 MySQL的下载 11](#_Toc3588636)

[6.2.2 MySQL的安装与使用 12](#_Toc3588637)

[6.2.3 Navicat for MySQL的配置与使用 12](#_Toc3588638)

[七.参考资料及博客 15](#_Toc3588639)

# 一.项目介绍

爬虫，是一个很奇妙的存在。它能代替你在网络上进行网页浏览，搜索，甚至可以模拟一切你在网页上的操作。使我们不必再一笔一划地记录数据并录入设备，只要一键就可以导入我们想要的数据，给我们的生活带来了十分优质的体验。本项目是利用对于初学者门槛较低的Beautifulsoup4库，实现对于豆瓣网站TOP250电影相关资料的爬取。

# 二.依赖库介绍

## 2.1 Urllib3

urllib 库是Python内置的HTTP请求库。urllib 模块提供的上层接口使访问www 和 ftp 上的数据就像访问本地文件一样。Urllib3库适用于3.x的python版本程序。

urllib有以下几种模块：urllib.request(请求模块),urllib.error(异常处理模块),urllib.parse (url 解析模块),urllib.robotparser robots.txt (解析模块)。本项目需要用到的是urllib.request模块。

## 2.2 Random

random库是用于产生随机数的Python标准库，在很多地方都能派的上用场。

## 2.3 Time

time库同样作为python中最为常用的库之一，可以用于记录时间等常用信息。除此之外还可以延长程序等待执行的时间。

## 2.4 Beautifulsoup4

BeautifulSoup是python的一个HTML或XML的解析库，我们可以用它来方便的从网页中提取数据，它拥有强大的API和多样的解析方式。它提供一些简单的方法和python式函数，用于浏览，搜索和修改解析树，它是一个工具箱，通过解析文档为用户提供需要抓取的数据。

## 2.5 Xlwt

Xlwt是python库中专门对excel文件进行操作的库。可以使用代码写入数据及改变其单元格样式，但Xlwt库目前较老（最终更新2017.8.22），推荐使用XlsxWriter库（最终更新2019.2.23），几乎所有在Xlwt库中的常用操作在XlsxWriter库都可以实现。本项目将使用Xlwt库进行操作。

# 三.环境配置

## 3.1操作系统

Windows10家庭版

版本号1803

## 3.2 python版本

本项目使用Python 3.7.0版本执行

## 3.3编辑器

Spyder (3.3.3版本)

由于自己不小心将PyQt5包升到最新版本(5.12)导致了Spyder打不开的问题，这里强调一下PyQt5包的版本为5.9.2。

## 3.4 库的安装

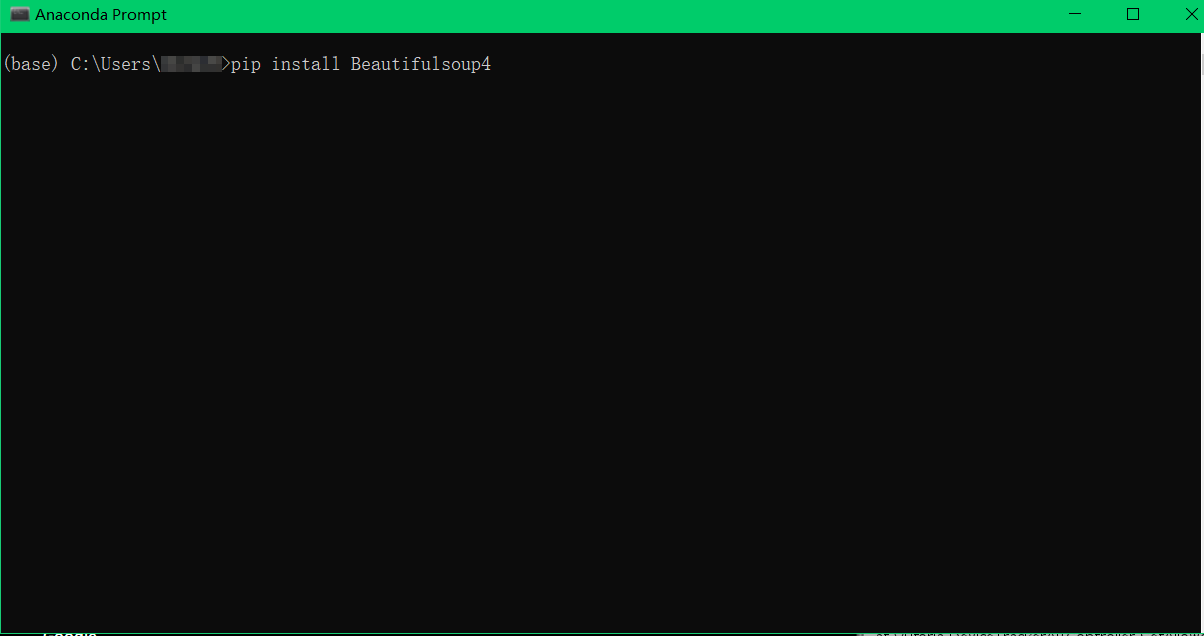
#### 3.4.1 python自带库

Time,random是python自带库，无需另外安装。

#### 3.4.2 Beautifulsoup4库安装

首先打开Anaconda Prompt

输入 pip install Beautifulsoup4 并按下回车，如下图所示。



等待命令行出现’Successfully installed Beautifulsoup4-4.7.1’字样（截至2019年3月12日）表示其已安装成功。

#### 3.4.3 urllib3库安装

操作方法与Beautifulsoup4库安装几乎相同

首先打开Anaconda Prompt

输入’pip install Beautifulsoup4 ’并按下回车，等待其安装完成即可。

#### 3.4.4 xlwt库的安装

首先打开Anaconda Prompt

输入’pip install xlwt ’并按下回车，等待其安装完成即可。

# 四. 执行网络爬虫前的准备

## 4.1目标网页介绍

目标网页网址：https://movie.douban.com/top250

豆瓣网算是已知的反爬虫机制较为人性化的网站，即使是直接不加修饰地去爬取信息也不会担心IP地址被永久封禁，因此较适合入门的程序员进行爬虫代码的练习。

## 4.2 爬虫前的其他代码详解

#### 4.2.1获取目标页面

由于豆瓣TOP250网页信息分别显示于10个网页上，如果要使用爬虫爬取其所有信息，需要构造全部十个网站的网址。具体规律如下（图为第二，三页网址）

捕获b

捕获c

以此类推，根据其规律可以构造第二到第十页的网址进行爬虫的准备。



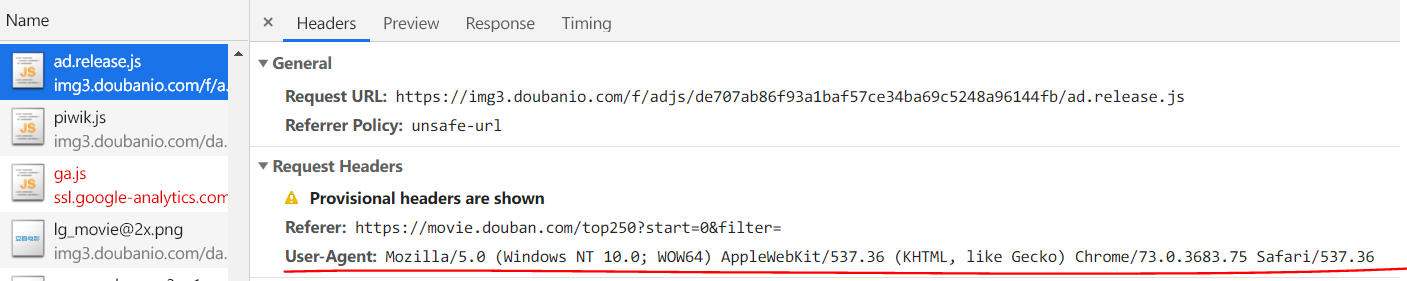
本部分使用了time.sleep()函数（括号内代表程序秒数，可以为浮点数），目的是防止过于高频率地访问目标网页而导致获取信息失败，经过多次测试发现，休眠一秒以上基本可以稳定获取目标信息，而设定0.5秒时会爬取失败。

#### 4.2.2反爬虫机制的对策

##### 4.2.2.1寻找Http请求头信息（这里以豆瓣网TOP250为例）

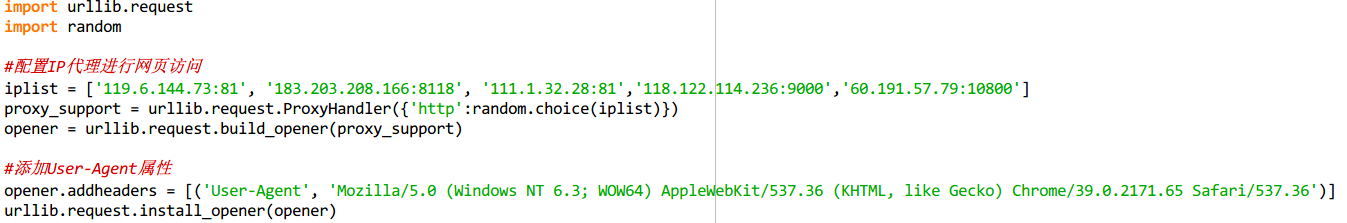
请求头则是其浏览器向网页服务器发出的授权信息，也是爬虫项目中不可缺少的一部分。

在豆瓣网TOP250页面按下F12（本人使用的浏览器是Chorme 73.0.3683.75\_32位版本），在第一栏选择Network（显示网页相关get以及post信息），在Name一栏中随意选择一个文件，在Headers中最下方即可找到你的网页浏览器的请求头信息。



##### 4.2.2.2 配置代理ip地址并添加Headers属性

IP地址，给用户上网使用的网际协议的设备的数字标签，可以看作你访问网页时本地网络的身份证明，过高频率访问其网页时，目标网页可能会将你的身份识别为爬虫从而将你的ip地址列入黑名单，所以为了避免这种情况的发生，我们需要足够用的“身份证明”来伪装不同的身份访问网页。而添加Headers属性则是用于模拟浏览器进行网页访问的必要环节。



本部分使用了random库的random.choice()方法来随机选择IP地址来达到不被轻易被目标网站服务器识别的目的。

##### 4.2.2.3 time库的使用

time库中的time.sleep()（括号内代表程序暂停秒数，可以为浮点数）方法可以使程序执行到此位置时暂停一段时间，是爬虫代码中的常见部分，对于对时间较为敏感的目标网页效果明显。其目的是防止过于高频率地访问目标网页而导致获取信息失败，经过多次测试发现，休眠一秒以上基本可以稳定获取目标信息，而设定0.5时秒会爬取失败。本项目由于每次构造网页时都会传递一次网址参数，所以对构造网页的代码进行改写，如下图所示。



# 五. 爬虫本体部分的代码讲解

## 5.1 爬取免费代理网站并获取完整IP地址

由于固定且有限的ip地址（代理服务器）经常会被关闭或者被过度使用，风险较大。

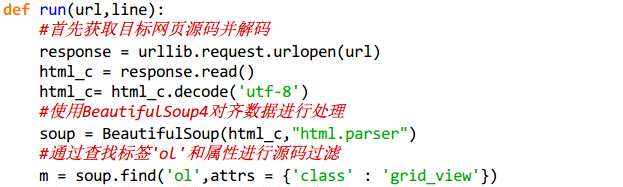
所以可以先使用一个简单的爬虫将其页面的代理IP爬取出来作为正式爬虫项目的ip地址库。

(在开启网页时建议不要将网页拓展代理打开(如Proxy SwitchyOmega),可能会导致网页源代码获取失败)。



## 5.2 使用 beautifulsoup4库 对豆瓣网站TOP250电影相关信息进行爬取

首先收集源代码并解码为’utf-8’的编码格式，再使用beautifulsoup4库进行信息处理，然后找出包涵所有需要获取的信息的目标标签。部分代码如下。



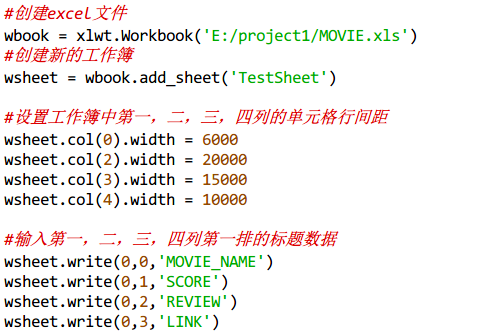
然后对高度重复的标签进行遍历访问，分别获取不同的信息并存储于一个列表中。而图片则存储为.JPG文件进行本地存储。形式参数中line是为了记录爬虫在源码中爬取的数据集的地点，同时也可以用于对爬取的图片文件名的拼接构造。url则是在正式爬取前构造完成的10个网页地址，在此部分代码中正式被用于爬取目标信息。部分代码如下。(注：保存文件时如果路径中含有不存在的文件夹会显示报错，所以应该在写文件前将文件夹创建完成)



# 六.爬取信息之后数据存储

## 6.1 将数据存储进入Excel文件中

只是模拟浏览其网页信息不是我们的最终目的，我们需要将其爬取的信息进行存储以便于项目组或其他人员进行数据分析。将信息存入Excel文件中是其常用方法之一。本项目使用Xlwt库进行数据的存入操作。部分代码如下。

xlwt

这时新的Excel文件已经创建完毕，单元格样式也已设置完毕，可以导入数据了。本项目使用的是在爬取数据同时直接将数据写入Excel文件的方法。



此时即可将其爬取的数据存入Excel文件中。

## 6.2 将文件信息存入数据库

本项目使用的数据库管理系统是MySQL，由于其体积小、速度快、总体拥有成本低，尤其是开放源码这一特点而被广为使用。

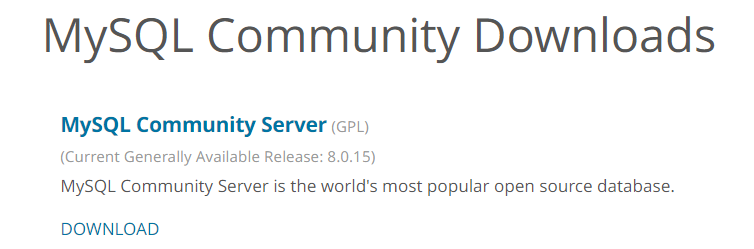
MySQL官网下载地址：https://www.mysql.com/downloads/

### **6.2.1MySQL的下载**

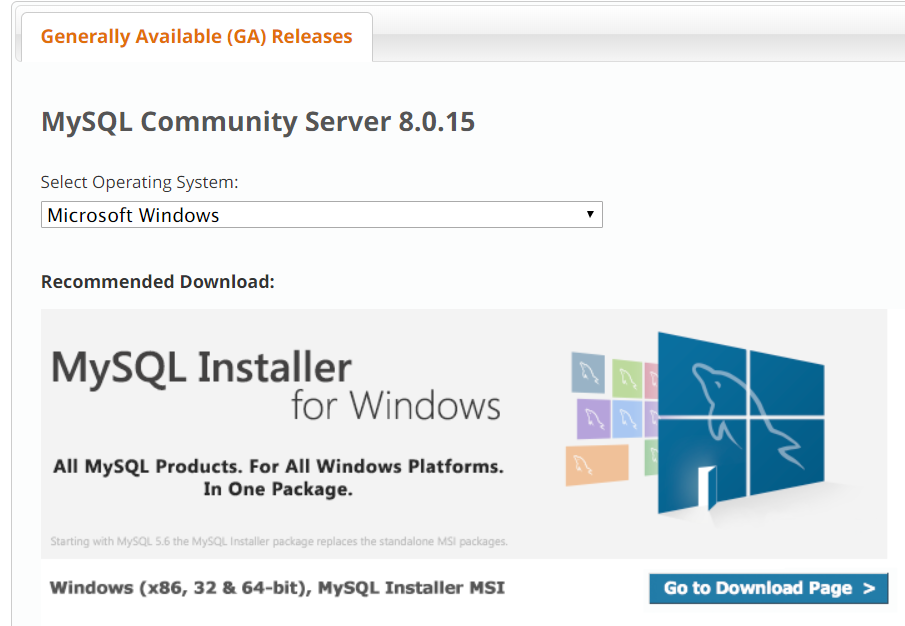
进入其官网下载地址，找到并点击Community(GPL) Downloads。如下图所示。



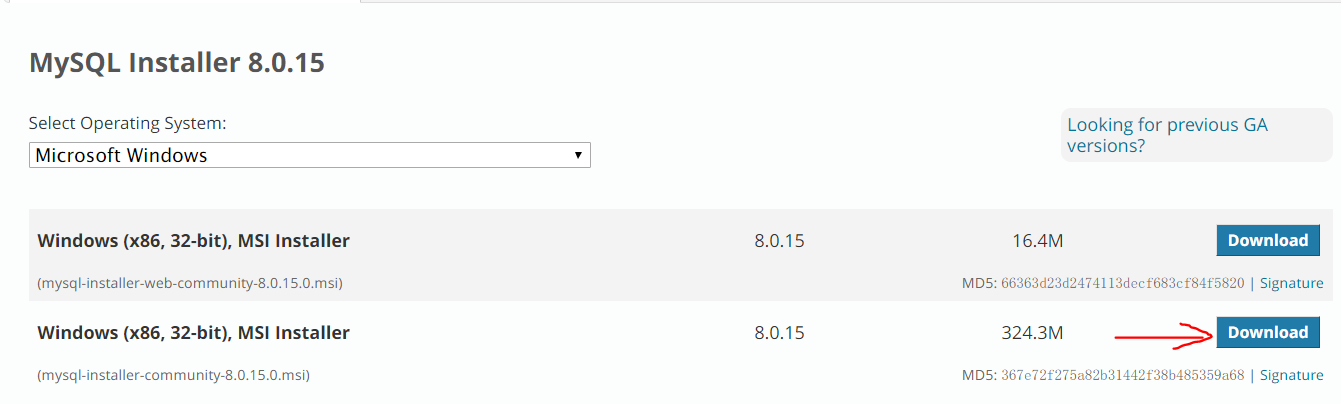
找到MySQL Community Server，点击下面的DOWNLOAD，进入下载页



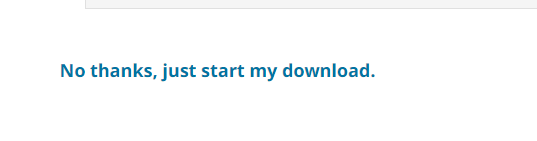
到了真正的下载页面，选择平台，选择Windows系统的版本，点击Download



此处建议下载下面mysql-installer-community-8.0.15.0.msi的版本，上面较小的文件是在线安装版本。



最后点击No thanks, just start my download.直接下载即可



### 6.2.2 MySQL的安装与使用

使用了离线安装包(如6.2.1所述)进行安装的话，并不需要进行特殊设置，一路同意并next即可。

到了installer界面显示Accounts and Roles 的界面中设置服务器密码，在Windows services Name中设置服务器名字。

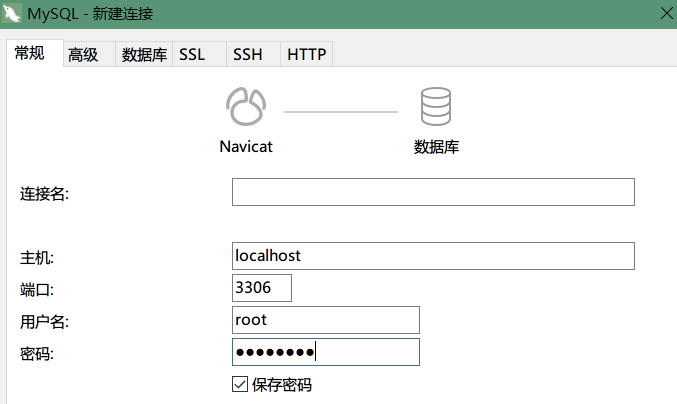
### 6.2.3 Navicat for MySQL的配置与使用

官方下载地址：<https://www.navicat.com.cn/download/navicat-for-mysql>

可以根据自己的情况进行选择。

Navicat for MySQL 是管理和开发 MySQL 或 MariaDB 的理想解决方案。它是一套单一的应用程序，能同时连接 MySQL 和 MariaDB 数据库。这套全面的前端工具为数据库管理、开发和维护提供了一款直观而强大的图形界面。此软件即使不使用数据库语言也能轻松管理MySQL数据库，并且轻松地通过图形界面将数据显示出来。安装过程同样没有特别需要注意的地方，这里不再赘述。

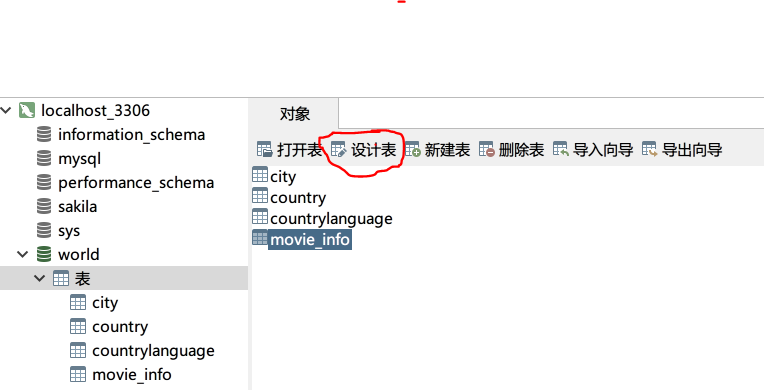
安装后打开界面，打开’连接’，输入你在MySQL中设置的root密码。如下图所示



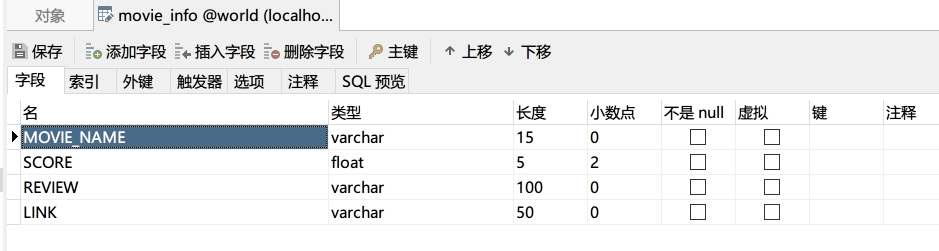
连接成功后操作界面右方会显示你的MySQL相关信息。

本项目使用Navicat for mysql实现对数据进行从Excel文件向MySQL数据库的转存操作。

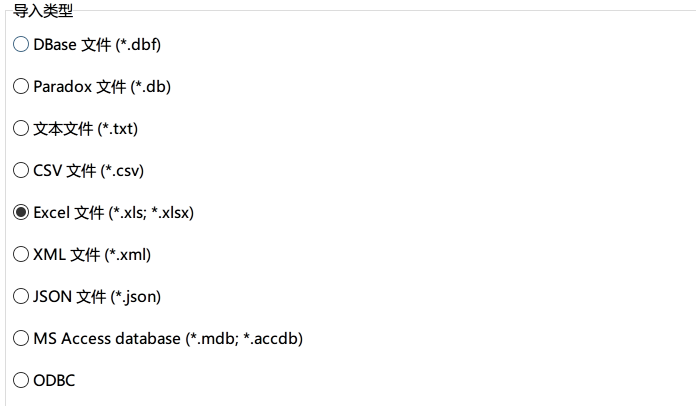
首先新建表并设置名字，然后点击设计表进行传输数据前的准备。



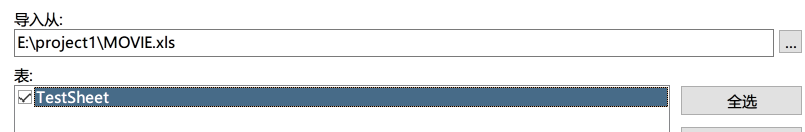
之后对需要存入的数据进行类型设置(varchar为可变长度字符串，float为浮点型数据)。设置完成后保存。



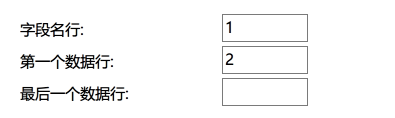
现在正式开始传输数据，先点击导入向导，选择Excel文件。



之后导入你想要导入的数据源(Excel文件)并选择对应工作簿。



接下来可以设置其导入数据的格式。将Excel文件中从第二排数据开始进行导入。

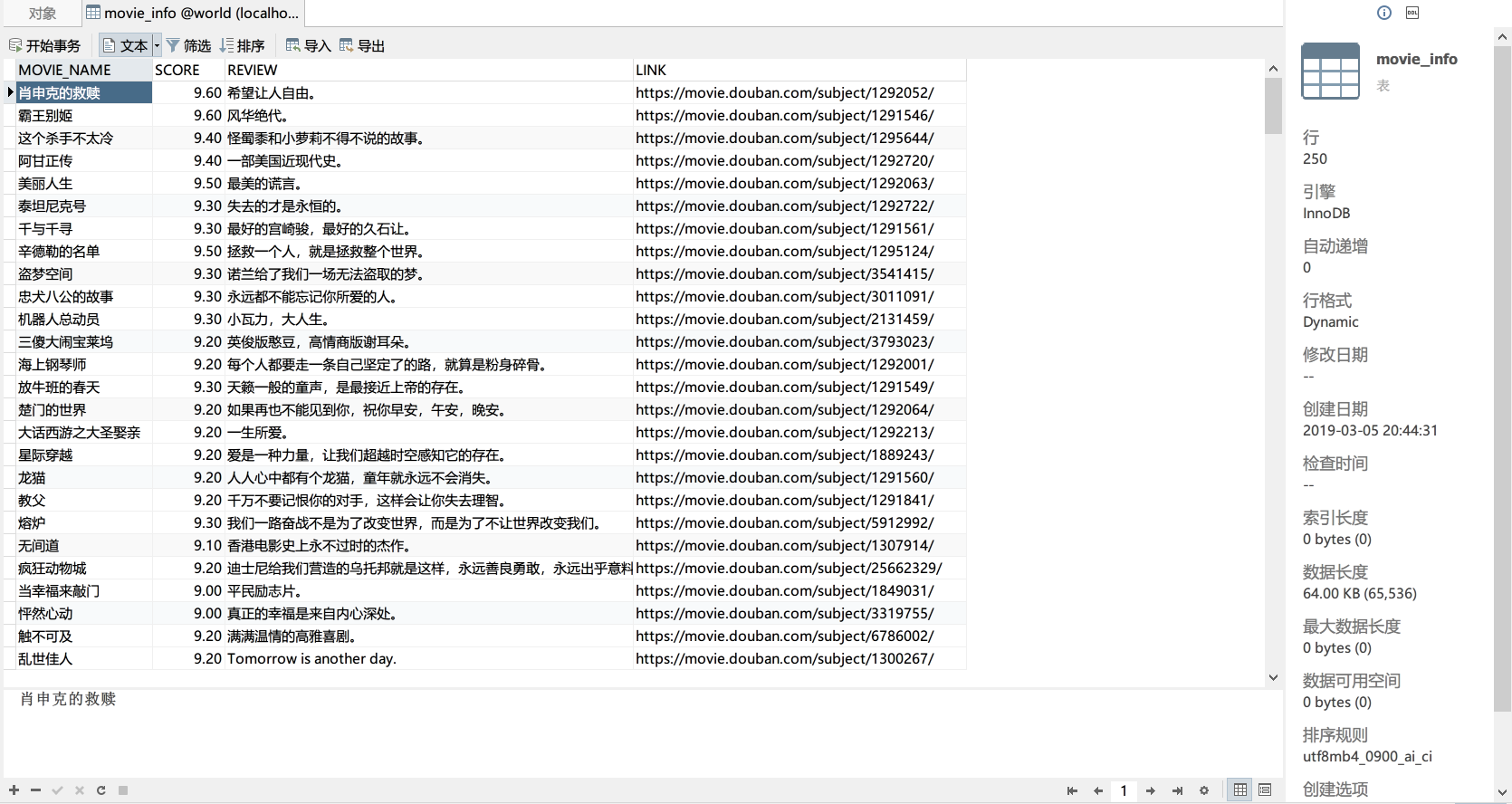


选择目标表并对应字段，导入模式建议选择复制，接下来即可导入。

下载mysql7

看到此字样即表示导入成功。

重新打开导入完成的表，其信息已完全被录入。



# 七.参考资料及博客

安装MySQL与Navicat的详细过程：<http://www.cnblogs.com/sshoub/p/4321640.html>

关于BeautifulSoup4库的使用教程：<https://www.cnblogs.com/zhangxinqi/p/9218395.html>

本项目源代码的地址：

<https://github.com/OGT968/Web-Crawler-for-DoubanTOP250-by-Using-Beautifulsoup4>