Scrapy框架爬取的数据存放到MySQL数据库中

# Python3.x 安装PyMySQL

## 安装MySQL

略，就是一般的**MySQL**。

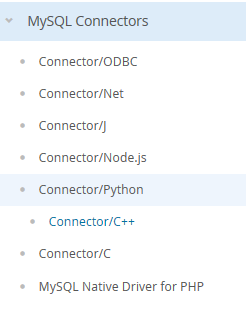
## 利用pip安装PyMySQL库：

**pip install PyMySQL**

注意：在Python2.x中使用的是MySQLdb，在Python3.x中使用PyMySQL。

## 安装 mysql-connector

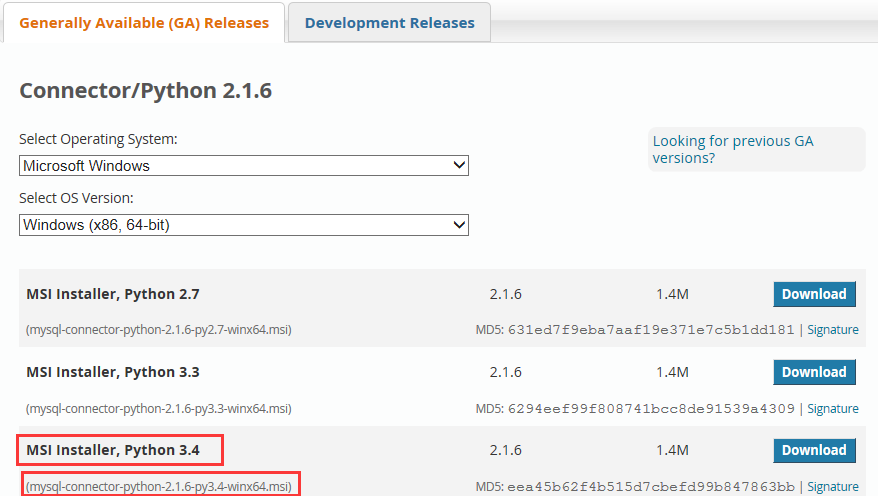
任何编程语言，都需要驱动包连接MySQL，去<https://dev.mysql.com/downloads/connector/> 下载相应的驱动包即可。



### .msi文件安装

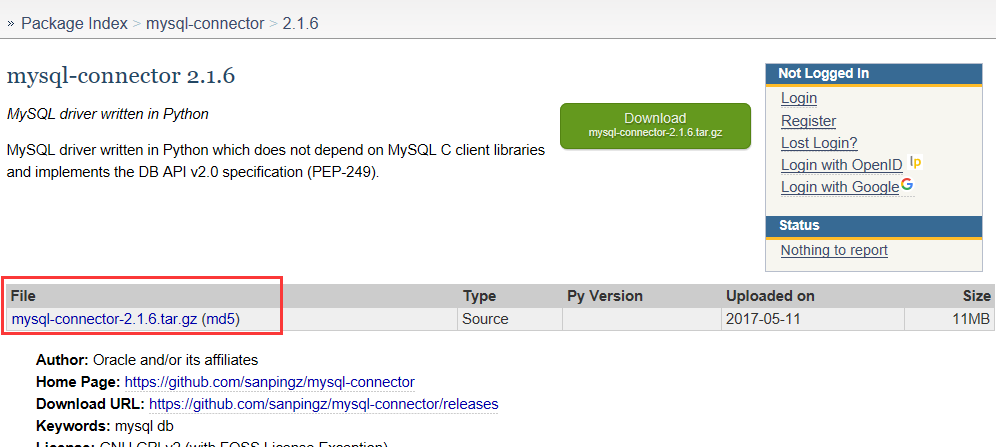
**去MySQL的网站上下载.msi文件安装。**

[**https://dev.mysql.com/downloads/connector/python/**](https://dev.mysql.com/downloads/connector/python/)



### 压缩包解压安装

<https://pypi.python.org/pypi/mysql-connector>

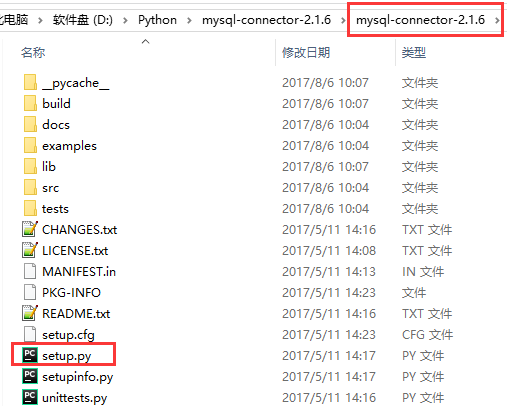


解压后，打开命令行，利用命令安装：

进入到解压后的目录：

**python setup.py build**

**python setup.py install**



# Python3.x中使用PyMySQL连接MySQL数据库

## 需要导入两个包

**import pymysql**

**import pymysql.cursors**

## 获取连接connector的方法

**利用pymql的connect方法**。

### 利用直接传入相应的参数

myConnector = pymysql.connect(host= **'127.0.0.1'**,port= 3306,user=**'root'**,password=**'zhao12'**, db=**'xicidaili'**,charset=**'utf8mb4'**,

cursorclass=pymysql.cursors.DictCursor)

### 把信息先入一个字典dict中，然后传入pymysql.connect(\*\*dict)方法中获取连接。

config = {  
 **'host'**: **'127.0.0.1'**,  
 **'port'**: 3306,  
 **'user'**: **'root'**,  
 **'password'**: **'zhao12'**,  
 **'db'**: **'xicidaili'**,  
 **'charset'**: **'utf8mb4'**,  
 **'cursorclass'**: pymysql.cursors.DictCursor  
 }  
 **myConnector = pymysql.connect(\*\*config)**

### 将配置信息写入settings.py文件中，并获取

**从settings.py文件中获取方法：config = spider.settings.get(‘CONF\_MYSQL’)，此时返回的config是一个字典类型<class 'dict'>。**

import pymysql  
BOT\_NAME = 'mysqlLianxi'  
SPIDER\_MODULES = ['mysqlLianxi.spiders']  
NEWSPIDER\_MODULE = 'mysqlLianxi.spiders'  
LOG\_FILE = './log.txt'  
CONF\_MYSQL = {  
 'host': '127.0.0.1',  
 'port': 3306,  
 'user': 'root',  
 'password': 'zhao12',  
 'db': 'mysqllianxi',  
 'charset': 'utf8mb4',  
 'cursorclass': **pymysql.cursors.DictCursor** # 需要导入pymysql，不能写成 字符串，即不能加引号  
}

CONF\_MYSQL = spider.settings.get('CONF\_MYSQL')  
myLogger.info(type(CONF\_MYSQL))#<class 'dict'> 为字典类型  
myConn = pymysql.connect(\*\*CONF\_MYSQL)#一定加两个\*\*



注意：**最后一个cursorclass属性值一定不要写成字符串，否则报错。**

## 增删改查的方法

### 编写sql语句

sql = ('insert into univList(univRank,univName,univProName,univScore) values(%s,%s,%s,%s)')

### 调用cursor的execute方法执行

这里**myConnector**对象类型为<class '**pymysql.connections.Connection**'>；

cur的类型为<class '**pymysql.cursors.DictCursor**'>。

**DictCursor 具有execute方法**；

Connection具有**rollback**回滚方法和**commit**提交方法。

**pipeline.py 文件示例**：

import pymysql  
import logging  
myLogger = logging.getLogger('赵红')  
class MysqllianxiPipeline(object):  
 def process\_item(self, item, spider):  
 sql = ('insert into univList(univRank,univName,univProName,univScore) values(%s,%s,%s,%s)')  
 data = (item['univRank'],item['univName'],item['univProName'],item['univScore'])  
 CONF\_MYSQL = spider.settings.get('CONF\_MYSQL')  
 myLogger.info(type(CONF\_MYSQL))#<class 'dict'>  
 myConn = pymysql.connect(\*\*CONF\_MYSQL)  
 with myConn.cursor() as cursor:  
 try:  
 cursor.execute(sql, data)  
 except Exception as e:  
 myConn.rollback()  
 myLogger.error(str(e))  
 finally:  
 myConn.commit()  
 cursor.close()  
 myConn.close()  
 return item

# 中国大学排名示例：存放到Mysql数据库中

## 目标分析

**目标网站**：<http://www.zuihaodaxue.com/zuihaodaxuepaiming2016.html>



将**红色部分**内容存放到数据库中，有后面的是通过JavaScript返回的，所以，不获取。

利用Scrapy爬虫框架，需要利用pymysql模块。

## 创建数据库及表格

数据库：create database mysqllianxi default charset='utf8';

表格：create table **univList2**(id int(10) primary key auto\_increment,univRank int(10),univName varchar(20),univProName varchar(20),univScore varchar(20)) default charset='utf8';

## scrapy编程：

### items.py编程：确定所需要的字段信息

from scrapy import Item,Field  
class MysqllianxiItem(Item):  
 # define the fields for your item here like:  
 univRank = Field()  
 univName = Field()  
 univProName = Field()  
 univScore = Field()

### settings.py ：配置scrapy

import pymysql  
BOT\_NAME = 'mysqlLianxi'  
SPIDER\_MODULES = ['mysqlLianxi.spiders']  
NEWSPIDER\_MODULE = 'mysqlLianxi.spiders'  
LOG\_FILE = './log.txt'  
CONF\_MYSQL = {  
 'host': '127.0.0.1',  
 'port': 3306,  
 'user': 'root',  
 'password': 'zhao12',  
 'db': 'mysqllianxi',  
 'charset': 'utf8mb4',  
 'cursorclass': pymysql.cursors.DictCursor#需要导入pymql  
}  
# Obey robots.txt rules  
ROBOTSTXT\_OBEY = False  
ITEM\_PIPELINES = {  
 'mysqlLianxi.pipelines.MysqllianxiPipeline': 300,  
}

### pipeline.py:对数据结果操作

import pymysql  
import logging  
myLogger = logging.getLogger('赵红')  
class MysqllianxiPipeline(object):  
 def process\_item(self, item, spider):  
 sql = ('insert into univList(univRank,univName,univProName,univScore) values(%s,%s,%s,%s)')  
 data = (item['univRank'],item['univName'],item['univProName'],item['univScore'])  
 CONF\_MYSQL = spider.settings.get('CONF\_MYSQL')  
 myLogger.info(type(CONF\_MYSQL))#<class 'dict'>  
 myConn = pymysql.connect(\*\*CONF\_MYSQL)  
 with myConn.cursor() as cursor:  
 try:  
 cursor.execute(sql, data)  
 except Exception as e:  
 myConn.rollback()  
 myLogger.error(str(e))  
 finally:  
 myConn.commit()  
 cursor.close()  
 myConn.close()  
 return item

### 编写spider：mysql.py

import scrapy  
from mysqlLianxi.items import MysqllianxiItem  
import logging  
mylog = logging.getLogger('赵红')  
  
class MysqlSpider(scrapy.Spider):  
 name = 'mysql'  
 start\_urls = ['http://www.zuihaodaxue.com/zuihaodaxuepaiming2016.html']  
  
 def parse(self, response):  
 trs = response.xpath('//tbody[@class="hidden\_zhpm"]/tr')  
 for tr in trs:  
 tdList = tr.xpath('td')  
 univRank = tdList[0].xpath('text()').extract()[0]  
 univName = tdList[1].xpath('div').xpath('text()').extract()[0]  
 univProName = tdList[2].xpath('text()').extract()[0]  
 univScore = tdList[3].xpath('text()').extract()[0]  
 item = MysqllianxiItem()  
 item['univRank'] = univRank  
 item['univName'] = univName  
 item['univProName'] = univProName  
 item['univScore'] = univScore  
 yield item

**结果**：

