# 大神博客：

python爬虫大数据：<https://ask.hellobi.com/blog/seng/category/517>

爬虫系统：<http://mp.weixin.qq.com/s?__biz=MzI2MzQ1NTEwNg==&mid=2247483698&idx=2&sn=cffef134a5470a79567f86d403f22cb7&chksm=eabae9dfddcd60c93b485d1c503030a0563c9fc8706d3fa01e41d5ffd14e998f2bb827d8ad12&mpshare=1&scene=23&srcid=0419WJScIAz2xJJb1T15Irul#rd>

# 环境搭建

## 1 python安装

必须：运行 yum install sqlite-devel（方便后面安装sqlite数据库）

yum install zlib-devel （方便安装setuptools）

yum install gcc libffi-devel python-devel openssl-devel

安装后需要重新编译环境 make&make install

**1.刚开始centos可能会缺少gcc等组件，先安装组件**

yum groupinstall "Development Tools"

**2.下载源码，解压后进入目录**

#下载地址[https://www.python.org](https://www.python.org/" \t "http://www.cnblogs.com/longxiang92/p/_blank)

#解压

tar xvJf \*\*\*.tar.xz#进入目录

cd Python-3.5\*

**3.安装python**

#创建目录

mkdir /usr/python3.5

#运行配置 指定要安装到/usr/python3.5目录下 并共享

./configure --prefix=/usr/python3.5 --enable-shared  
  
make#安装

make install

**4.创建Python3.5软连接**

\*\*在/usr/bin中有python、python2、python2.7三个文件依次指向后者（即都是python2.7.5）

\*\*不建议依照其他网上的方法把默认的python修改为指向3.5版本，这样的话将无法使用firewall防火墙，而且可能还有其他有影响

\*\*建议另外创建一个python3的软连接

cd /usr/bin  
#创建python3.5版本的软连接

ln -s /usr/python3.5/bin/python3 /usr/bin/python3

#查看python版本  
python3 -V

**5.如果运行查看版本失败**

错误信息类似这样 ：error while loading shared libraries: xxxxxxxx: cannot open shared object file: No such file or directory（xxxx为文件名）

进入安装文件夹（解压出来的目录）

#拷贝生成的文件到lib

cp xxxx（libpython3.5m.so.1.0） /usr/local/lib

#修改ldconfig文件

vim /etc/ld.so.conf

#新加一行 /usr/local/lib

运行命令ldconfig

Ldconfig

6、设置别名方便使用

alias py=python3

## 2安装setuptools（easy\_install）

**方法一（推荐）：**

wget --no-check-certificate  https://pypi.python.org/packages/source/s/setuptools/setuptools-19.6.tar.gz#md5=c607dd118eae682c44ed146367a17e26

tar -zxvf setuptools-19.6.tar.gz

cd setuptools-19.6

python3 setup.py build

python3 setup.py install

**方法二（有时不太行）：**

最新版下载地址：<https://pypi.python.org/pypi/ez_setup>

tar -zxvf setuptools-3.6.tar.gz

cd setuptools-3.6

py setup.py install

卸载：easy\_install -mxN networkx

## 3 pip安装

从官网下载最新pip：<https://pypi.python.org/pypi/pip>

tar zvxf 1.5.5 #解压文件

cd pip-1.5.5/

python3 setup.py install

# vim /etc/profile

在文档最后，添加:

export PATH="/usr/local/python2.7/bin:$PATH"

pip查看已安装的包 pip show --files paramiko 或者pip list

更新pip ：python -m pip install --upgrade pip

pip检查哪些包需要更新 pip list --outdated

pip升级包 pip install --upgrade Django

pip卸载包 pip uninstall Django

## 4 安装scrapy

Pip安装 parsel, w3lib, cryptography, pyOpenSSL 如：pip install parsel

安装 lxml 和 twisted 网址：<http://www.lfd.uci.edu/~gohlke/pythonlibs/>

windos下：pip install E:\demo\Twisted-16.4.1-cp35-cp35m-win\_amd64.whl

Twisted要安装zope.interface,可以从下面地址下载

[zope.interface](https://pypi.python.org/packages/source/z/zope.interface/zope.interface-4.0.1.tar.gz):https://pypi.python.org/pypi/zope.interface

[twisted:](http://twistedmatrix.com/Releases/Twisted/12.1/Twisted-12.1.0.tar.bz2)http://twistedmatrix.com/Releases/Twisted

最后安装scrapy：pip install scrapy

打开端口：

iptables -I INPUT -p tcp --doprt 6800 -j ACCEPT

## 5 安装scrapyd

安装scprayd，网址：[https://github.com/scrapy/scrapyd](https://github.com/scrapy/scrapyd" \t "http://blog.csdn.net/xiaoquantouer/article/details/_blank)

安装scrapyd-client，网址：[https://github.com/scrapy/scrapyd-client](https://github.com/scrapy/scrapyd-client" \t "http://blog.csdn.net/xiaoquantouer/article/details/_blank)

用easy\_install安装scrapyd为最新版

Linux下配置文件路径：

/usr/python3.5.3/lib/python3.5/site-packages/scrapyd-1.2.0a1-py3.5.egg/scrapyd

python安装目录下的 Lib\site-packages\scrapyd\default\_scrapyd.conf (Windows)

mkdir eggs

mkdir logs

mkdir items

mkdir dbs

sudo mkdir /etc/scrapyd/

chown -R cdz /etc/scrapyd/

cd /etc/scrapyd/

vi scrapyd.conf 配置如下：  
\* **以下是几个比较重要的参数**：

http\_port 监听端口

bind\_address 监听IP地址

max\_proc 最大进程数（为0指开启等同于cpu的数目的进程）

max\_proc\_per\_cpu 每个cpu开启的进程数，默认：4

debug 是否开启调试模式

eggs\_dir 存储工程egg文件的目录

dbs\_dir 存储工程数据库的目录

logs\_dir 存储日志的目录（要关闭日志则让这个配置为空

logs\_filename = {spider}-{Y}{m}{d}.log

### 调度爬虫

1、添加版本信息

curl http://localhost:6800/addversion.json -F project=myproject -F     version=r23 -F egg=@myproject.egg

2、调度spider执行

curl http://127.0.0.1:6800/schedule.json -d project=bricCrawler -d   spider=somespider

3、取消spider的执行

curl http://localhost:6800/cancel.json -d project=myproject -d      job=6487ec79947edab326d6db28a2d8651

4、列出工程

curl [http://localhost:6800/listprojects.json](http://localhost:6800/listprojects.json？Project=bricCrawler)

5、列出job

curl http://localhost:6800/listjobs.json？Project=bricCrawler

## 6 定时框架apschedule

easy\_install apscheduler

简单例子：<http://stackoverflow.com/questions/29765039/how-to-use-apscheduler-with-scrapy>

详细操作：<http://debugo.com/apscheduler/>

## 7 持久层ORM框架：

tornado+sqlalchemy

Sqlalchemy使用: <http://www.jianshu.com/p/e6bba189fcbd>

更详细的方法详解：<http://www.jb51.net/article/49789.htm>

scrapy结合sqlalchemy+redis：<http://www.code123.cc/1434.html>

sqlchemy详细使用经验：<https://www.keakon.net/2012/12/03/SQLAlchemy%E4%BD%BF%E7%94%A8%E7%BB%8F%E9%AA%8C>

Sqlalchemy链接sqlserver：需要安装pymssql

# pyodbc  
engine = create\_engine('mssql+pyodbc://scott:tiger@mydsn')  
  
# pymssql  
engine = create\_engine('mssql+pymssql://scott:tiger@hostname:port/dbname')

## 8 PDF解析工具

### 1.pdfminer3k

pip install pdfminer3k

网址：<https://pypi.python.org/pypi/pdfminer3k/>

详解：<http://www.thinksaas.cn/topics/0/686/686331.html>

<http://m.php.cn/article/359112.html>

Pdf解密：

PDF是一个加密的PDF，所以无法解析 ，但是如果直接打开PDF却是可以的并没有要求输密码什么的，原因是这个PDF虽然是加过密的，但密码是空，所以就出现了这样的问题。

解决这个的问题的办法是通过qpdf命令来解密文件（要确保已经安装了qpdf），要想在python中调用该命令只需使用call即可：

1 from subprocess import call2 call('qpdf --password=%s --decrypt %s %s' %('', file\_path, new\_file\_path), shell=True)

### 2 pdftables

pip install git+https://github.com/pdftables/python-pdftables-api.git

### 3 tabula-py

只支持单页

试试tabula，读取pdf后可转为pandas dataframe进行后续处理，也可直接输出csv文件。  
python版本的项目主页，有示例代码：[https://github.com/chezou/tabula-py](https://link.zhihu.com/?target=https://github.com/chezou/tabula-py" \t "_blank)  
安装：pip install tabula-py

如果包含中文内容需要修改编码格式：  
pd.read\_csv("example.csv", encoding="GB18030")  
参考：<https://www.zhihu.com/question/22016184>

(ps下次提问请先查看万能的stackoverflow：  
<http://stackoverflow.com/questions/17217194/extracting-table-contents-from-a-collection-of-pdf-files/26110587#26110587>

4

<https://blog.pdftron.com/2014/05/25/a-simple-example-of-converting-pdf-to-html/>

# Scrapy使用

## 常用技巧

### Unicode编码：

data=json.dumps(intoDB)  
logging.info(**'传送数据{}'**.format(data.encode(**'raw-unicode-escape'**)))

### 时间减一天：

yesterday=self.today-datetime.timedelta(days=1)

例：

now = datetime.datetime.now()-datetime.timedelta(days=1)

day1 = now.strftime(**'%Y%m%d'**)

### Xpath中使用正则：

xpath(  
 **r'//div[@class="xh-nei-text"]/div/div[contains(strong/text(), "金乡：")]/following-sibling::div[2]/text()'**,  
 namespaces={**"re"**: **"http://exslt.org/regular-expressions"**})

### requests处理gb2312编码的网页：

response = requests.get(**'http://www.51garlic.com/jiage/kwsj1/'**)  
response.encoding = **'gb2312'**text = response.text

### Python处理gb2312编码：

1.

import chardet

......

        content\_type = chardet.detect(html\_content)

        #print(content\_type['encoding'])

        if content\_type['encoding'] != "UTF-8":

            html\_content = html\_content.decode(content\_type['encoding'])

        html\_content = html\_content.encode("utf-8")

        open(filename,"wb").write(html\_content)

2.

page=response.body.decode(response.encoding)

res=Selector(text=page)

res.xpath()

### 时间格式化为带中文：

today = datetime.datetime.now().strftime(**'%Y{y}%m{m}%d{d}'**)

.format(y=**'年'**, m=**'月'**, d=**'日'**)

### Scrapy处理json

json\_body=json.loads(response.body\_as\_unicode())

### 中文转url编码：

**from** urllib.parse **import** quote

quote(keyword)