1. 安装

使用scrapy前提

1.安装zlib

* Python 2.6 or 2.7
* Tar zxvf Python-2.7.5.tgz

### 安装zope.interface

1. 安装easy\_install

Cd setuptools-0.6c11

Python setup.py install

1. 安装Scrapy

Easy\_Install Scrapy

1. 安装OpenSSL

easy\_install OpenSSL

1. 安装lxml

easy\_install lxml

1. 安装w3lib

easy\_install w3lib

1. 安装Twisted

easy\_install Twisted

1. 升级开发环境  
   yum install xinetd telnet-server samba samba-common samba-client make gcc tftp-server patch gettext libtool gcc-c++ nfs-utils autoconf automake libtool gperf bison flex gtk-doc xorg-x11-util-macros
2. 安装gtk

下载pygtk-2.24.0.tar.bz

tar jxvf pygtk-2.24.0.tar.bz2

cd pygtk-2.24.0

下载依赖库[**http://ftp.gnome.org/pub/gnome/sources/pygobject/2.28/pygobject-2.28.6.tar.xz**](http://ftp.gnome.org/pub/gnome/sources/pygobject/2.28/pygobject-2.28.6.tar.xz)

安装  
xz -d pygobject-2.28.6.tar.xz

tar xvf pygobject-2.28.6.tar

cd pygobject-2.28.6

补丁[**http://www.linuxfromscratch.org/patches/blfs/svn/pygobject-2.28.6-introspection-1.patch**](http://www.linuxfromscratch.org/patches/blfs/svn/pygobject-2.28.6-introspection-1.patch)

cd pygobject-2.28.6

patch -p1 < ../pygobject-2.28.6-introspection-1.patch

下载依赖[**http://ftp.gnome.org/pub/gnome/sources/glib/2.34/glib-2.34.3.tar.xz**](http://ftp.gnome.org/pub/gnome/sources/glib/2.34/glib-2.34.3.tar.xz)

./configure --prefix=/usr/local

下载依赖libffi-3.0.13

[**ftp://sourceware.org/pub/libffi/libffi-3.0.13.tar.gz**](ftp://sourceware.org/pub/libffi/libffi-3.0.13.tar.gz)

补丁

[**http://www.linuxfromscratch.org/patches/blfs/svn/libffi-3.0.13-includedir-1.patch**](http://www.linuxfromscratch.org/patches/blfs/svn/libffi-3.0.13-includedir-1.patch)

cd **libffi-3.0.13**

**patch -Np1 -i ../libffi-3.0.13-includedir-1.patch**

**./configure --prefix=/usr --disable-static**

Make

**make install**

**设置**PKG\_CONFIG\_PATH环境变量

export PKG\_CONFIG\_PATH=/usr/lib/pkgconfig

下载PCRE-8.32

[**http://downloads.sourceforge.net/pcre/pcre-8.32.tar.bz2**](http://downloads.sourceforge.net/pcre/pcre-8.32.tar.bz2)

**./configure --prefix=/usr --docdir=/usr/share/doc/pcre-8.32 --enable-utf --enable-unicode-properties --enable-pcregrep-libz --enable-pcregrep-libbz2 --disable-static**

Make

Make install

下载依赖pkg-config

[**http://pkgconfig.freedesktop.org/releases/pkg-config-0.28.tar.gz**](http://pkgconfig.freedesktop.org/releases/pkg-config-0.28.tar.gz)

**./configure --prefix=/usr \**

**--docdir=/usr/share/doc/pkg-config-0.28 \**

**--with-internal-glib \**

**--disable-host-tool**

Make

Make check

Make install

下载依赖[**http://ftp.gnome.org/pub/gnome/sources/glib/2.34/glib-2.34.3.tar.xz**](http://ftp.gnome.org/pub/gnome/sources/glib/2.34/glib-2.34.3.tar.xz)

Cd glib-2.34.3

./configure --prefix=/usr **--with-pcre=system**

Make

Make install

# Linux(RedHat,Centos)上scrapy详尽安装笔记

2012-05-27 09:25:13cnblogs.com-sharpstill-点击数:326

**原创作者:博客园sharpstill,转载请注明**

Scrapy是一款非常成熟的爬虫框架，可以抓取网页数据并抽取结构化数据，目前已经有很多企业用于生产环境。对于它的更多介绍，可以查阅相关资料（官方网站：www.scrapy.org）。

我们根据官网提供的安装指南，来一步步安装，主要参考了http://doc.scrapy.org/en/latest/intro/install.html页面的介绍：

view plain

1. Requirements?0?9
2. Python 2.5, 2.6, 2.7 (3.x is not yet supported)
3. Twisted 2.5.0, 8.0 or above (Windows users: you’ll need to install Zope.Interface and maybe pywin32 because of this Twisted bug)
4. w3lib
5. lxml or libxml2 (if using libxml2, version 2.6.28 or above is highly recommended)
6. simplejson (not required if using Python 2.6 or above)
7. pyopenssl (for HTTPS support. Optional, but highly recommended)

下面记录一下从安装Python到安装scrapy的过程，最后，通过执行命令进行抓取数据来验证我们所做的安装配置工作。

## 准备工作

操作系统：RHEL 5  
Python版本：Python-2.7.2  
zope.interface版本：zope.interface-3.8.0  
Twisted版本：Twisted-11.1.0  
libxml2版本：libxml2-2.7.4.tar.gz  
w3lib版本：w3lib-1.0  
Scrapy版本：Scrapy-0.14.0.2841

## 安装配置

### 1、安装zlib

首先检查一下你的系统中是否已经安装zlib，该库是一个与数据压缩相关的工具包，scrapy框架依赖于该工具包。我使用的RHEL 5系统，查看是否安装：

view plain

1. [root@localhost scrapy]# rpm -qa zlib
2. zlib-1.2.3-3

我的系统已经默认安装了，安装的话，可以跳过该步骤。如果没有安装的话，可以到http://www.zlib.net/上下载，并进行安装。假如下载的是zlib-1.2.5.tar.gz，安装命令如下所示：view plain

1. [root@localhost scrapy]# tar -xvzf zlib-1.2.5.tar.gz
2. [root@localhost zlib-1.2.5]# cd zlib-1.2.5
3. [root@localhost zlib-1.2.5]# make
4. [root@localhost zlib-1.2.5]# make install

### 在我的centos上安装zlib的步骤是yum search zlib,然后yum install zlib-devel

### 2、安装Python

我的系统中已经安装的Python 2.4，根据官网要求和建议，我选择了Python-2.7.2，下载地址如下所示：

http://www.python.org/download/（需要代理）  
http://www.python.org/ftp/python/2.7.2/Python-2.7.2.tgz

我下载了Python的源代码，重新编译后，进行安装，过程如下所示：

view plain

1. [root@localhost scrapy]# tar -zvxf Python-2.7.2.tgz
2. [root@localhost scrapy]# cd Python-2.7.2
3. [root@localhost Python-2.7.2]# ./configure
4. [root@localhost Python-2.7.2]# make
5. [root@localhost Python-2.7.2]# make install

默认情况下，Python程序被安装到/usr/local/lib/python2.7。

如果你的系统中没有安装过Python，此时通过命令行执行一下：

view plain

1. [root@localhost scrapy]# python
2. Python 2.7.2 (default, Dec  5 2011, 22:04:07)
3. [GCC 4.1.1 20070105 (Red Hat 4.1.1-52)] on linux2
4. Type "help", "copyright", "credits" or "license" for more information.
5. >>>

表示最新安装的Python已经可以使用了。

如果你的系统中还有其他版本的Python，例如我的系统中2.4版本的，所以要做一个符号链接：

view plain

1. [root@localhost python2.7]# mv /usr/bin/python /usr/bin/python.bak
2. [root@localhost python2.7]# ln -s /usr/local/bin/python /usr/bin/python

这样操作以后，在执行python，就生效了。

### 3、安装setuptools

这里主要是安装一个用来管理Python模块的工具，如果已经安装就跳过该步骤。如果你需要安装，可以参考下面的链接：

wget http://pypi.python.org/packages/2.7/s/setuptools/setuptools-0.6c11-py2.7.egg#md5=fe1f997bc722265116870bc7919059ea  
chmod +x setuptools-0.6c11-py2.7.egg  
./ setuptools-0.6c11-py2.7.egg

不过，在安装Python-2.7.2以后，可以看到Python的解压缩包里面有一个setup.py脚本，使用这个脚本可以安装Python一些相关的模块，执行命令：

view plain

1. [root@localhost Python-2.7.2]# python setup.py install

安装执行后，相关Python模块被安装到目录/usr/local/lib/python2.7/site-packages下。

### 4、安装zope.interface

下载地址如下所示：

http://pypi.python.org/pypi/zope.interface/3.8.0  
http://pypi.python.org/packages/source/z/zope.interface/zope.interface-3.8.0.tar.gz#md5=8ab837320b4532774c9c89f030d2a389

安装过程如下所示：

view plain

1. [root@localhost scrapy]$ tar -xvzf zope.interface-3.8.0.tar.gz
2. [root@localhost scrapy]$ cd zope.interface-3.8.0
3. [root@localhost zope.interface-3.8.0]$ python setup.py build
4. [root@localhost zope.interface-3.8.0]$ python setup.py install

安装完成后，可以在/usr/local/lib/python2.7/site-packages下面看到zope和zope.interface-3.8.0-py2.7.egg-info。

### 5、安装Twisted

下载地址如下所示：

http://twistedmatrix.com/trac/  
http://pypi.python.org/packages/source/T/Twisted/Twisted-11.1.0.tar.bz2#md5=972f3497e6e19318c741bf2900ffe31c  
 替换为：http://twistedmatrix.com/Releases/Twisted/12.0/Twisted-12.0.0.tar.bz2 **注意:scrapy 0.14配合 twisted 11.1.0，目前发现bug，爬着爬着会出现不能爬行的情况。可以考虑降级twisted至11.0.0 或 升级scrapy至0.15 或升级twisted至12.0.0解决。**

安装过程如下所示：

view plain

1. [root@localhost scrapy]# bzip2 -d Twisted-11.1.0.tar.bz2
2. [root@localhost scrapy]# tar -xvf Twisted-11.1.0.tar
3. [root@localhost scrapy]# cd Twisted-11.1.0
4. [root@localhost Twisted-11.1.0]# python setup.py install

安装完成后，可以在/usr/local/lib/python2.7/site-packages下面看到twisted和Twisted-11.1.0-py2.7.egg-info。

### 6、安装w3lib

下载地址如下所示：

http://pypi.python.org/pypi/w3lib  
http://pypi.python.org/packages/source/w/w3lib/w3lib-1.0.tar.gz#md5=f28aeb882f27a616e0fc43d01f4dcb21

安装过程如下所示：

view plain

1. [root@localhost scrapy]# tar -xvzf w3lib-1.0.tar.gz
2. [root@localhost scrapy]# cd w3lib-1.0
3. [root@localhost w3lib-1.0]# python setup.py install

安装完成后，可以在/usr/local/lib/python2.7/site-packages下面看到w3lib和w3lib-1.0-py2.7.egg-info。

### 7、安装libxml2

下载地址如下所示：

http://download.chinaunix.net/download.php?id=28497&ResourceID=6095  
http://download.chinaunix.net/down.php?id=28497&ResourceID=6095&site=1

或者，可以到网站http://xmlsoft.org上面找到相应版本的压缩包。

安装过程如下所示：

view plain

1. [root@localhost scrapy]# tar -xvzf libxml2-2.7.4.tar.gz
2. [root@localhost scrapy]# cd libxml2-2.7.4
3. [root@localhost libxml2-2.7.4]# ./configure
4. [root@localhost libxml2-2.7.4]# make
5. [root@localhost libxml2-2.7.4]# make install

### 8、安装pyOpenSSL

该步骤可选，对应的安装包下载地址为：

https://launchpad.net/pyopenssl

我在centos上安装这个pyOpenSSL死活提示有错，最后google到了https://bugs.launchpad.net/pyopenssl/+bug/845445这个解决方案。最后用easy\_install http://pypi.python.org/packages/source/p/pyOpenSSL/pyOpenSSL-0.12.tar.gz 一键安装解决。

如果这一步easy\_install仍然出错，则是因为没有安装openssl,在centos上使用yum install openssl-devel搞定后easy\_install

### 9、安装Scrapy

注：这一步也可以easy\_install scrapy一键搞定

下载地址如下所示：

http://scrapy.org/download/  
http://pypi.python.org/pypi/Scrapy  
http://pypi.python.org/packages/source/S/Scrapy/Scrapy-0.14.0.2841.tar.gz#md5=fe63c5606ca4c0772d937b51869be200

安装过程如下所示：

view plain

1. [root@localhost scrapy]# tar -xvzf Scrapy-0.14.0.2841.tar.gz
2. [root@localhost scrapy]# cd Scrapy-0.14.0.2841
3. [root@localhost Scrapy-0.14.0.2841]# python setup.py install

## 安装验证

经过上面的安装和配置过程，已经完成了Scrapy的安装，我们可以通过如下命令行来验证一下：

view plain

1. [root@localhost scrapy]# scrapy
2. Scrapy 0.14.0.2841 - no active project
4. Usage:
5. scrapy <command> [options] [args]
7. Available commands:
8. fetch         Fetch a URL using the Scrapy downloader
9. runspider     Run a self-contained spider (without creating a project)
10. settings      Get settings values
11. shell         Interactive scraping console
12. startproject  Create new project
13. version       Print Scrapy version
14. view          Open URL in browser, as seen by Scrapy
16. Use "scrapy <command> -h" to see more info about a command

上面提示信息，提供了一个fetch命令，这个命令抓取指定的网页，可以先看看fetch命令的帮助信息，如下所示：view plain

1. [root@localhost scrapy]# scrapy fetch --help
2. Usage
3. =====
4. scrapy fetch [options] <url>
6. Fetch a URL using the Scrapy downloader and print its content to stdout. You
7. may want to use --nolog to disable logging
9. Options
10. =======
11. --help, -h              show this help message and exit
12. --spider=SPIDER         use this spider
13. --headers               print response HTTP headers instead of body
15. Global Options
16. --------------
17. --logfile=FILE          log file. if omitted stderr will be used
18. --loglevel=LEVEL, -L LEVEL
19. log level (default: DEBUG)
20. --nolog                 disable logging completely
21. --profile=FILE          write python cProfile stats to FILE
22. --lsprof=FILE           write lsprof profiling stats to FILE
23. --pidfile=FILE          write process ID to FILE
24. --set=NAME=VALUE, -s NAME=VALUE
25. set/override setting (may be repeated)

根据命令提示，指定一个URL，执行后抓取一个网页的数据，如下所示：view plain

1. [root@localhost scrapy]# scrapy fetch http://doc.scrapy.org/en/latest/intro/install.html ;> install.html
2. 2011-12-05 23:40:04+0800 [scrapy] INFO: Scrapy 0.14.0.2841 started (bot: scrapybot)
3. 2011-12-05 23:40:04+0800 [scrapy] DEBUG: Enabled extensions: LogStats, TelnetConsole, CloseSpider, WebService, CoreStats, MemoryUsage, SpiderState
4. 2011-12-05 23:40:04+0800 [scrapy] DEBUG: Enabled downloader middlewares: HttpAuthMiddleware, DownloadTimeoutMiddleware, UserAgentMiddleware, RetryMiddleware, DefaultHeadersMiddleware, RedirectMiddleware, CookiesMiddleware, HttpCompressionMiddleware, ChunkedTransferMiddleware, DownloaderStats
5. 2011-12-05 23:40:04+0800 [scrapy] DEBUG: Enabled spider middlewares: HttpErrorMiddleware, OffsiteMiddleware, RefererMiddleware, UrlLengthMiddleware, DepthMiddleware
6. 2011-12-05 23:40:04+0800 [scrapy] DEBUG: Enabled item pipelines:
7. 2011-12-05 23:40:05+0800 [default] INFO: Spider opened
8. 2011-12-05 23:40:05+0800 [default] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
9. 2011-12-05 23:40:05+0800 [scrapy] DEBUG: Telnet console listening on 0.0.0.0:6023
10. 2011-12-05 23:40:05+0800 [scrapy] DEBUG: Web service listening on 0.0.0.0:6080
11. 2011-12-05 23:40:07+0800 [default] DEBUG: Crawled (200) <GET http://doc.scrapy.org/en/latest/intro/install.html> (referer: None)
12. 2011-12-05 23:40:07+0800 [default] INFO: Closing spider (finished)
13. 2011-12-05 23:40:07+0800 [default] INFO: Dumping spider stats:
14. {'downloader/request\_bytes': 227,
15. 'downloader/request\_count': 1,
16. 'downloader/request\_method\_count/GET': 1,
17. 'downloader/response\_bytes': 22676,
18. 'downloader/response\_count': 1,
19. 'downloader/response\_status\_count/200': 1,
20. 'finish\_reason': 'finished',
21. 'finish\_time': datetime.datetime(2011, 12, 5, 15, 40, 7, 918833),
22. 'scheduler/memory\_enqueued': 1,
23. 'start\_time': datetime.datetime(2011, 12, 5, 15, 40, 5, 5749)}
24. 2011-12-05 23:40:07+0800 [default] INFO: Spider closed (finished)
25. 2011-12-05 23:40:07+0800 [scrapy] INFO: Dumping global stats:
26. {'memusage/max': 17711104, 'memusage/startup': 17711104}
27. [root@localhost scrapy]# ll install.html
28. -rw-r--r-- 1 root root 22404 Dec  5 23:40 install.html
29. [root@localhost scrapy]#

可见，我们已经成功抓取了一个网页。

接下来，可以根据scrapy官网的指南来进一步应用scrapy框架，Tutorial链接页面为http://doc.scrapy.org/en/latest/intro/tutorial.html。

**注意：安装过程中如果发生gcc报错，则需要安装 libxml2和libxslt:**  
(1)如果在ubuntu下  
使用：$sudo apt-get install libxml2 libxml2-dev 安装libxml2 使用：$sudo apt-get install libxlst1 libxslt1-dev 安装libxslt   
(2)如果在centos下  
yum install libxml2  
yum install libxslt-devel  
可以先yum search一下

[的](http://www.th7.cn/e/tags/?tagname=%E7%9A%84) [官方网站](http://www.th7.cn/e/tags/?tagname=%E5%AE%98%E6%96%B9%E7%BD%91%E7%AB%99) [Linux](http://www.th7.cn/e/tags/?tagname=Linux) [网页](http://www.th7.cn/e/tags/?tagname=%E7%BD%91%E9%A1%B5) [资料](http://www.th7.cn/e/tags/?tagname=%E8%B5%84%E6%96%99)