

Scrapy介绍（二）

windows环境配置

Scrapy依赖包（也可到官网单独下载各文件安装）：

1.xml: pip install wheel

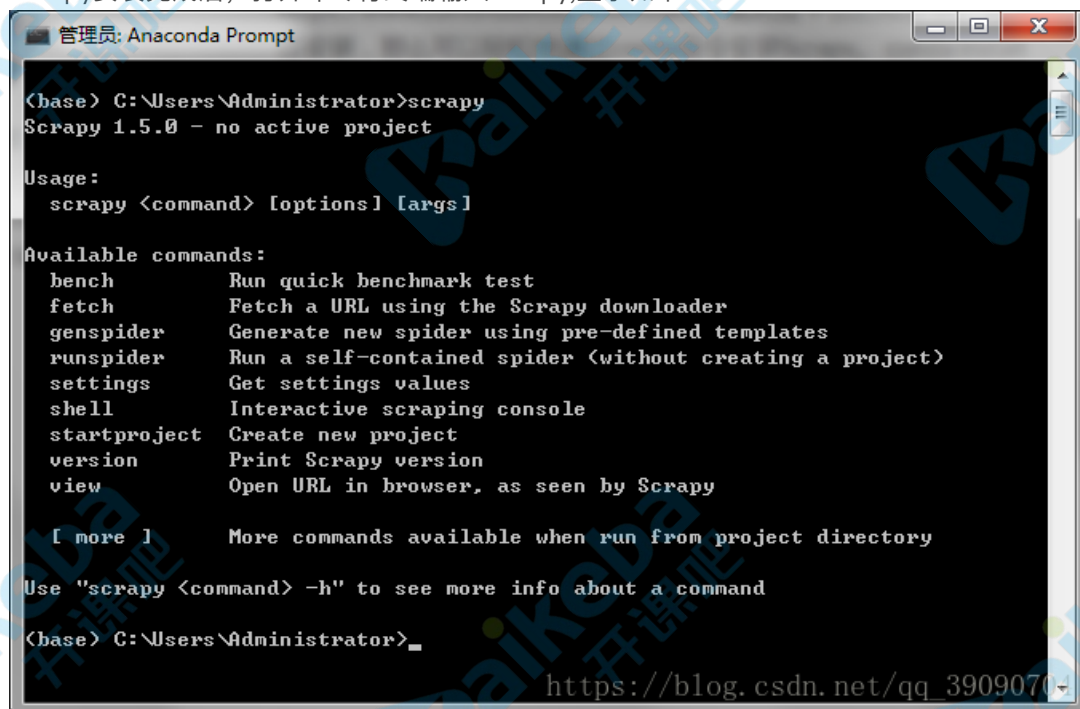
2.zope.interface: pip install zope.interface-4.3.3-cp35-cp35m-win_amd64.whl

3.pyOpenSSL: pip install pyOpenSSL

4.Twisted: pip install Twisted

5.Scrapy: pip install Scrapy

Scrapy安装完成后，打开命令行终端输入scrapy,显示如下：



```
管理员: Anaconda Prompt

(base) C:\Users\Administrator>scrapy
Scrapy 1.5.0 - no active project

Usage:
  scrapy <command> [options] [args]

Available commands:
  bench           Run quick benchmark test
  fetch           Fetch a URL using the Scrapy downloader
  genspider       Generate new spider using pre-defined templates
  runspider       Run a self-contained spider (without creating a project)
  settings        Get settings values
  shell           Interactive scraping console
  startproject    Create new project
  version         Print Scrapy version
  view           Open URL in browser, as seen by Scrapy

[ more ]         More commands available when run from project directory

Use "scrapy <command> -h" to see more info about a command

(base) C:\Users\Administrator>
```

https://blog.csdn.net/qq_3909070

创建项目

- 创建爬虫项目命令

scrapy startproject project_name

创建爬虫文件命令

scrapy genspider example example.com

文件目录如下：

```
D:\test>tree /F
卷 软件 的文件夹 PATH 列表
卷序列号为 58B6-0E53
D:.
```

```
├──project_dir
│   └── scrapy.cfg
│
├──project_name
│   ├──items.py
│   ├──middlewares.py
│   ├──pipelines.py
│   ├──settings.py
│   ├──__init__.py
│   ├──spiders
│   │   ├──__init__.py
│   │   ├──__pycache__
│   └──__pycache__
```

items.py: 定义爬虫程序的数据模型，类似于实体类。

middlewares.py: 爬虫中间件，负责调度。

pipelines.py: 管道文件，负责对spider返回数据的处理。

spiders目录 负责存放继承自scrapy的爬虫类

scrapy.cfg.scrapy 基础配置

init: 初始化文件

setting.py: 负责对整个爬虫的配置，内容如下

```
# -*- coding: utf-8 -*-

# Scrapy settings for baidu project
#
# For simplicity, this file contains only settings considered important or
# commonly used. You can find more settings consulting the documentation:
#
#     https://doc.scrapy.org/en/latest/topics/settings.html
#     https://doc.scrapy.org/en/latest/topics/downloader-middleware.html
#     https://doc.scrapy.org/en/latest/topics/spider-middleware.html

BOT_NAME = 'baidu'

# 爬虫所在地
SPIDER_MODULES = ['baidu.spiders']
```

```
NEWSPIDER_MODULE = 'baidu.spiders'
```

```
# Crawl responsibly by identifying yourself (and your website) on the user-agent  
#USER_AGENT = 'baidu (+http://www.yourdomain.com)'
```

```
# Obey robots.txt rules
```

```
# 遵守爬虫协议
```

```
ROBOTSTXT_OBEY = False
```

```
# Configure maximum concurrent requests performed by Scrapy (default: 16)
```

```
# 最大请求并发量 默认16
```

```
# CONCURRENT_REQUESTS = 32
```

```
# configure 配置 请求延迟
```

```
# Configure a delay for requests for the same website (default: 0)
```

```
# See https://doc.scrapy.org/en/latest/topics/settings.html#download-delay
```

```
# See also autothrottle settings and docs
```

```
#DOWNLOAD_DELAY = 3
```

```
# The download delay setting will honor only one of:
```

```
#CONCURRENT_REQUESTS_PER_DOMAIN = 16
```

```
#CONCURRENT_REQUESTS_PER_IP = 16
```

```
# Disable cookies (enabled by default)
```

```
# 是否使用cookie
```

```
#COOKIES_ENABLED = False
```

```
# Disable Telnet Console (enabled by default)
```

```
#TELNETCONSOLE_ENABLED = False
```

```
# Override the default request headers:
```

```
#DEFAULT_REQUEST_HEADERS = {
```

```
#   'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8',
```

```
#   'Accept-Language': 'en',
```

```
#}
```

```
# Enable or disable spider middlewares
```

```
# See https://doc.scrapy.org/en/latest/topics/spider-middleware.html
```

```
#SPIDER_MIDDLEWARES = {
```

```
#     值越小,优先级越高,优先级越高,越先执行
```

```
#     'baidu.middlewares.BaiduSpiderMiddleware': 543,
```

```
#}
```

```
# Enable or disable downloader middlewares
```

```
# See https://doc.scrapy.org/en/latest/topics/downloader-middleware.html
```

```
#DOWNLOADER_MIDDLEWARES = {
```

```
#     值越小,优先级越高,优先级越高,越先执行
```

```
#     'baidu.middlewares.BaiduDownloaderMiddleware': 543,
```

```
#}
```

```
# Enable or disable extensions 是否进行扩展
# See https://doc.scrapy.org/en/latest/topics/extensions.html
#EXTENSIONS = {
#     'scrapy.extensions.telnet.TelnetConsole': None,
#}

# Configure item pipelines
# See https://doc.scrapy.org/en/latest/topics/item-pipeline.html
ITEM_PIPELINES = {
    # 值越小,优先级越高,优先级越高,越先执行
    'baidu.pipelines.BaiduPipeline': 1,
}

# Enable and configure the AutoThrottle extension (disabled by default)
# See https://doc.scrapy.org/en/latest/topics/autothrottle.html
#AUTOTHROTTLE_ENABLED = True
# The initial download delay
#AUTOTHROTTLE_START_DELAY = 5
# The maximum download delay to be set in case of high latencies
#AUTOTHROTTLE_MAX_DELAY = 60
# The average number of requests Scrapy should be sending in parallel to
# each remote server
#AUTOTHROTTLE_TARGET_CONCURRENCY = 1.0
# Enable showing throttling stats for every response received:
#AUTOTHROTTLE_DEBUG = False

# Enable and configure HTTP caching (disabled by default)
# See https://doc.scrapy.org/en/latest/topics/downloader-middleware.html#httpcache-e-middleware-settings
#HTTPCACHE_ENABLED = True
#HTTPCACHE_EXPIRATION_SECS = 0
#HTTPCACHE_DIR = 'httpcache'
#HTTPCACHE_IGNORE_HTTP_CODES = []
#HTTPCACHE_STORAGE = 'scrapy.extensions.httpcache.FilesystemCacheStorage'
```