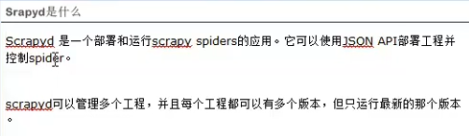
Scrapy的部署工具-Scrapyd

# scrapyd是什么?



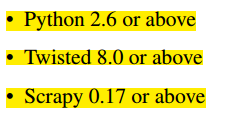
**Scrapyd is an application for deploying and running Scrapy spiders**. It enables you to deploy (upload) your projects and control their spiders **using a JSON API**.

Scrapyd can manage multiple projects and each project can have multiple versions uploaded, but only the latest one will be used for launching new spiders.

# 安装Scrapyd

## 前提条件Requirements

Scrapyd depends on the following libraries, but the installation process takes care of installing the missing ones:



## 安装命令：pip install scrapyd

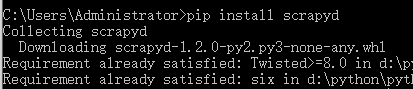
默认安装在 **D:\Python\Python36\Scripts** 目录下。（必要时，需要添加环境变量）

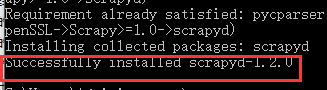
## 测试是否安装成功：启动scrapyd

在命令行下，执行scrapyd即可启动scrapyd服务。

## 示例：

### 安装scrapyd服务：





### 启用scrapyd 服务



### 浏览器查看：<http://localhost:6800/>

**可以访问即证明成功安装scrapyd。**

# 部署工程：Deploying your project：将工程部署到scrapyd上

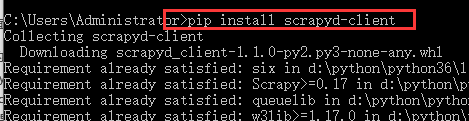
## 安装上传工具scrapyd-client

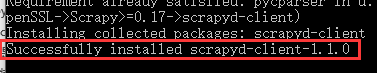


**安装方法： pip install scrapyd-client**

**默认安装在D:\Python\Python36\Scripts 目录下**。

（必要时，需要添加环境变量）





## 部署工程方法：scrapyd-deploy

**python** D:\Python\Python36\Scripts\**scrapyd-deploy** **target**–p **project**

### target：主机地址，如localhost

### project：工程名字，如myproject

### scrapyd-deploy不是一个exe可执行命令，在这里必须书写绝对路径；通过环境变量配置路径无效。

### 必须进入到scrapy工程的目录之下，才能成功部署。

### 可选参数：



### 官方文档介绍

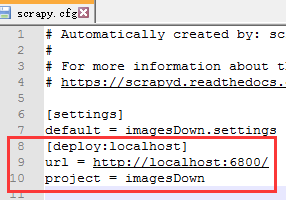
官方文档：Deploying your project involves eggifying it and uploading the egg to Scrapyd via the **addversion.json** endpoint. You can do this manually, **but the easiest way is to use the scrapyd-deploy tool provided by scrapyd-client** which will do it all for you.

## 部署错误：Unknown：localhost

### 解决方法：首先检查并修改待上传工程的scrapy.cfg文件

**一般默认情况下，不需要修改**。如果出现Unknown：localhost时，可以修改下。

**将[deploy]修改为[deploy:localhost]**



## 安装调度工具curl



<http://pan.baidu.com/s/1mgy9zVq>

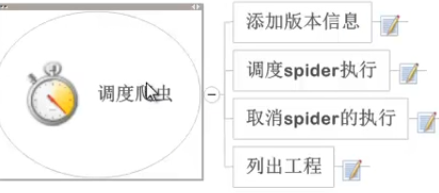
4lqh



安装好后测试：

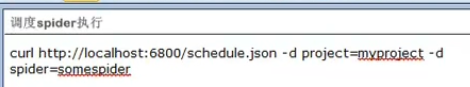


## 调度爬虫





curl [http://localhost:6800/**schedule.json**](http://localhost:6800/schedule.json) -d project=myproject -d spider=somespider



curl <http://localhost:6800/cancel.json> -d project=myproject -d job=\*\*\*\*\*\*\*\*



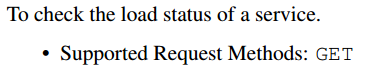
## 列出已经部署的所有工程

命令 ： **curl http://localhost:6800/listprojects.json**

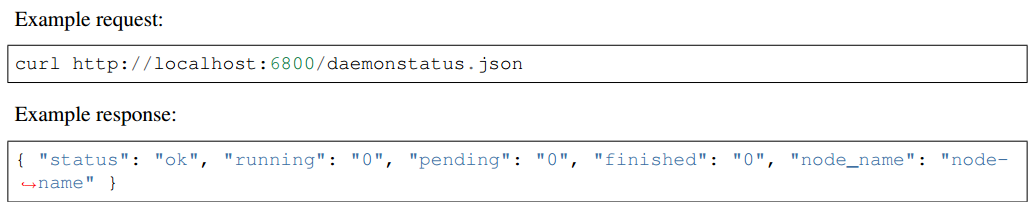


# API：scrapyd的Scrapyd JSON API

## daemonstatus.json

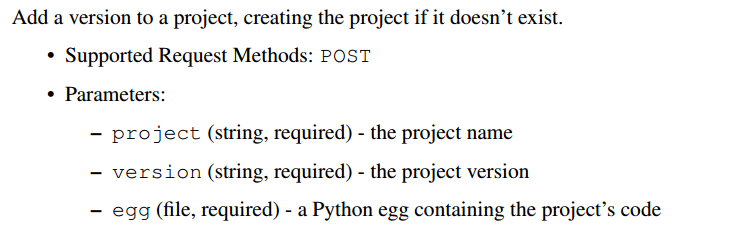


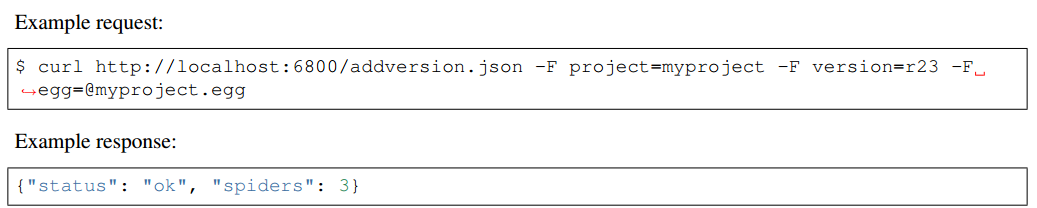
daemon：守护，守护进程。status状态。





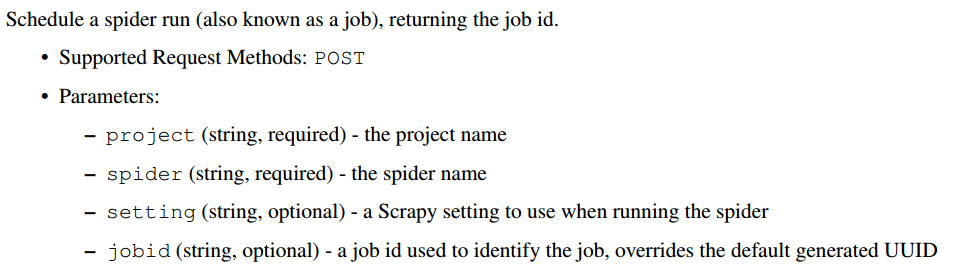
## addversion.json

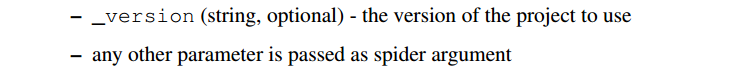


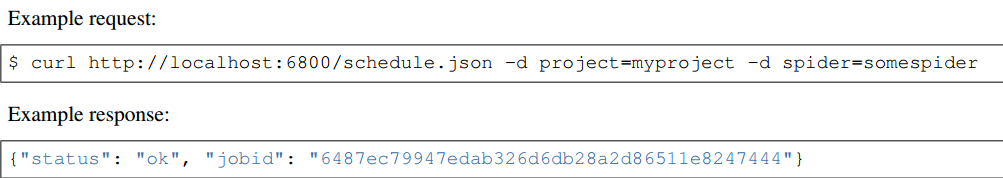


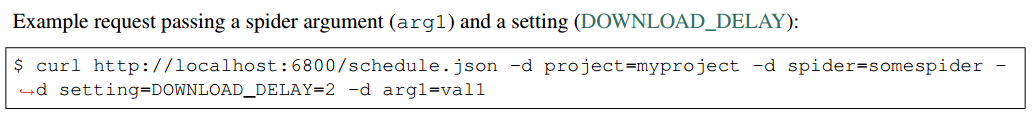
注意：The latest version for a project will be used by default whenever necessary.  **schedule.json and listspiders.json** allow you to explicitly set the desired project version.

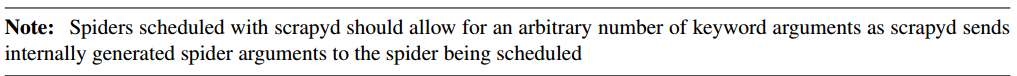
## schedule.json



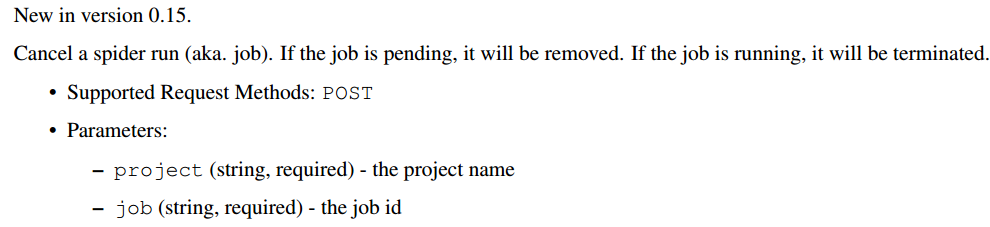


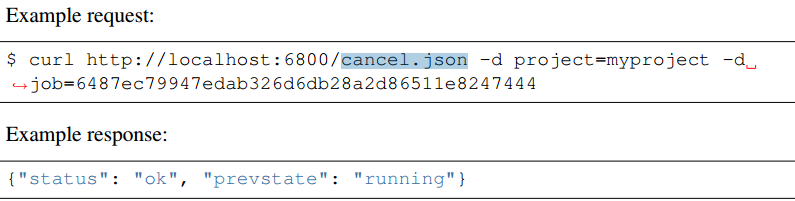




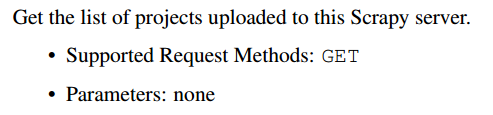


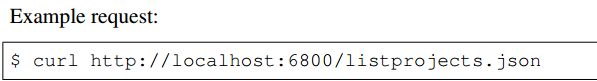
## cancel.json





## listprojects.json：列出所有的project



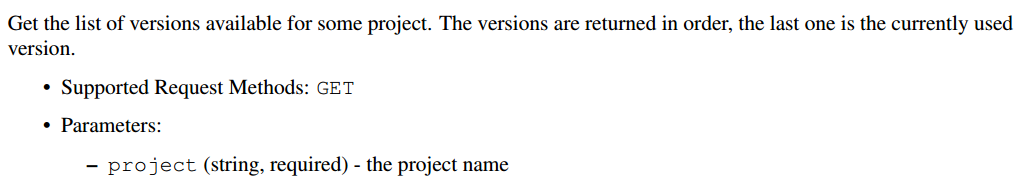


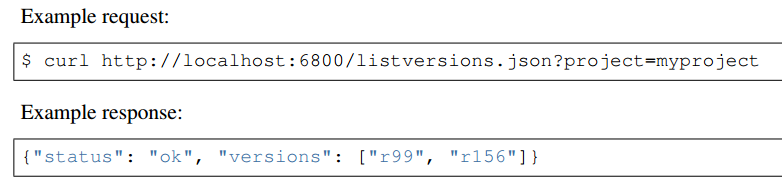






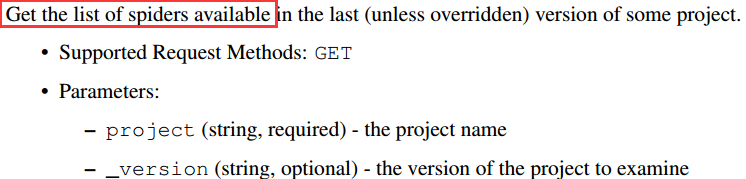
## listversions.json：列出一个Project的所有的version

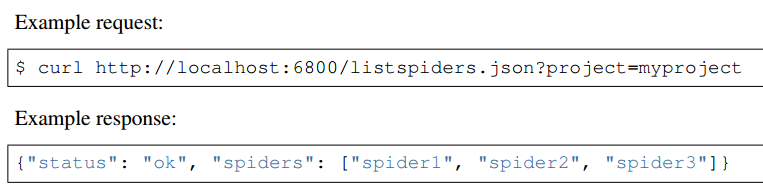






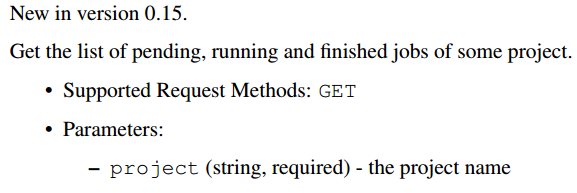
## listspiders.json

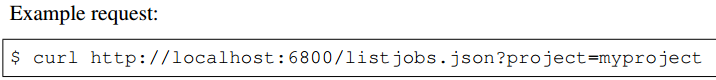


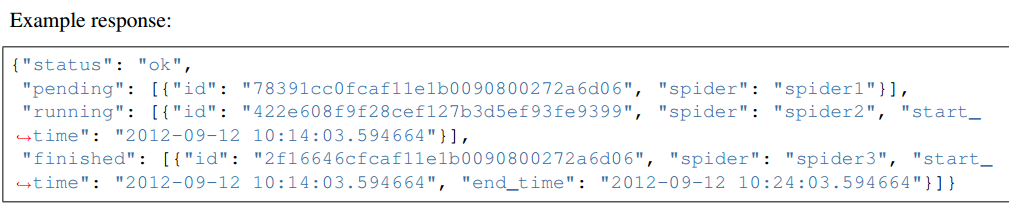




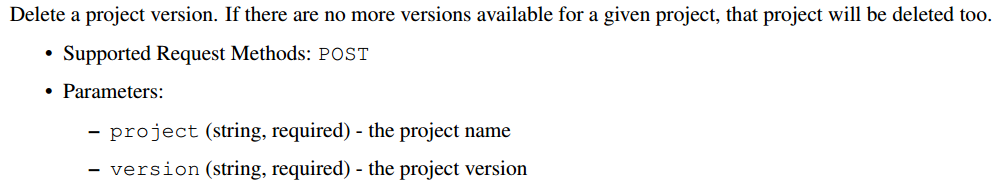
## listjobs.json

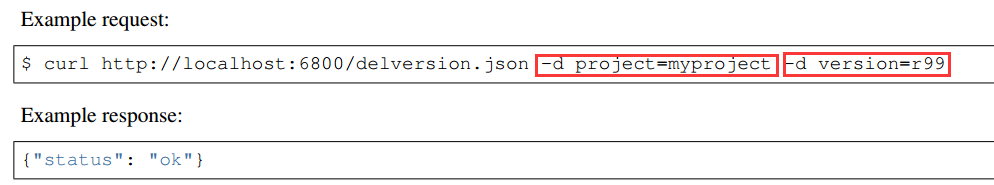




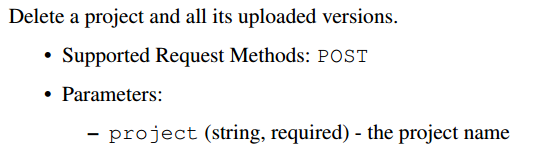


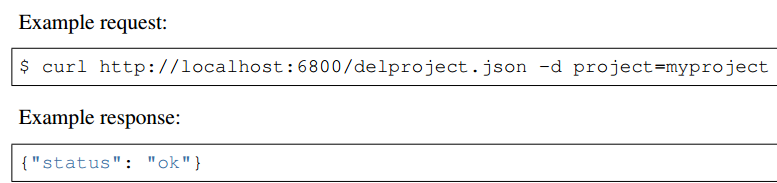
## delversion.json





## delproject.json

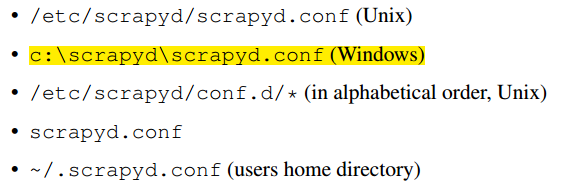




# scrapyd的配置文件：Configuration file

## 配置文件位置：scrapyd.conf或default\_scrapyd.conf

Scrapyd searches for **configuration files** in the following locations, and parses them in order with the latest one taking more priority:



## 具体配置选项介绍见官方文档

分成两大部分：scrapyd和services

## scrapyd.conf文件示例：

[scrapyd]

eggs\_dir = eggs

logs\_dir = logs

items\_dir =

jobs\_to\_keep = 5

dbs\_dir = dbs

max\_proc = 0

max\_proc\_per\_cpu = 4

finished\_to\_keep = 100

poll\_interval = 5.0

bind\_address = 127.0.0.1

http\_port = 6800

debug = off

runner = scrapyd.runner

application = scrapyd.app.application

launcher = scrapyd.launcher.Launcher

webroot = scrapyd.website.Root

[services]

schedule.json = scrapyd.webservice.Schedule

cancel.json = scrapyd.webservice.Cancel

addversion.json = scrapyd.webservice.AddVersion

listprojects.json = scrapyd.webservice.ListProjects

listversions.json = scrapyd.webservice.ListVersions

listspiders.json = scrapyd.webservice.ListSpiders

delproject.json = scrapyd.webservice.DeleteProject

delversion.json = scrapyd.webservice.DeleteVersion

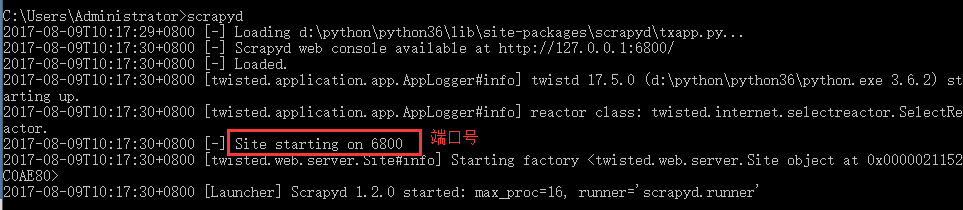
listjobs.json = scrapyd.webservice.ListJobs

daemonstatus.json = scrapyd.webservice.DaemonStatus

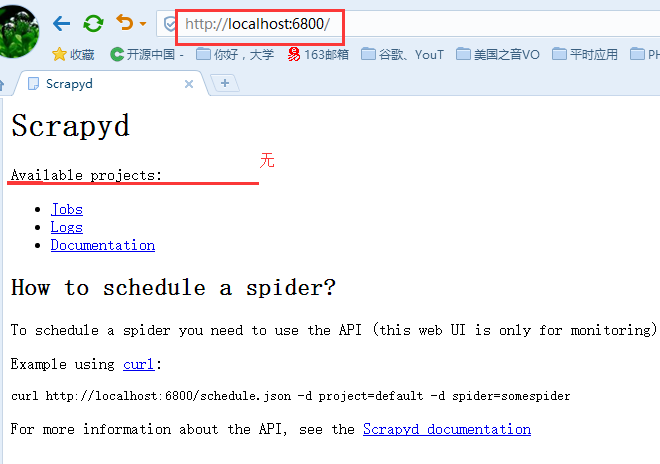
# 练习使用：

## 启动scrapyd

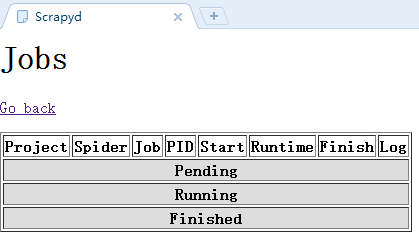
**在命令行任意目录下，输入scrapyd，即可启动scrapyd服务**。默认端口号是**6800**，如图所示：



**在浏览器中可以访问该服务**：



**Jobs页面可以查看正在执行的Job**。



## 部署工程imagesDown和imageDown2

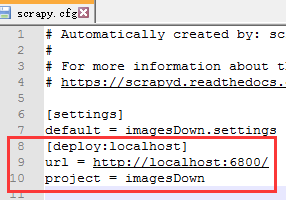
进入到**相应工程根目录**下，然后输入命令：

python D:/Python/Python36/Scripts/scrapyd-deploy **localhost** -p **imagesDown**

python D:/Python/Python36/Scripts/scrapyd-deploy **localhost** -p **imageDown2**

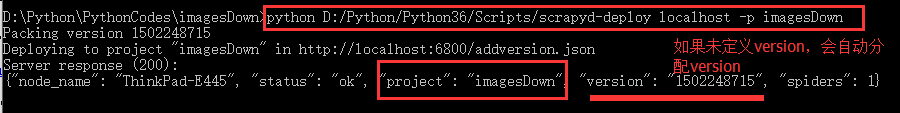
### 首先检查并修改待上传工程的scrapy.cfg文件

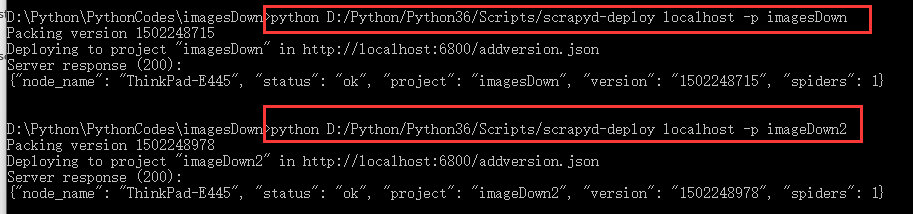
一般默认情况下，不需要修改。如果出现Unknown：localhost时，可以修改下。

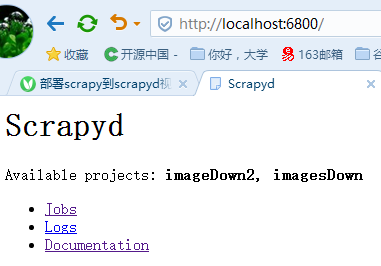


### 利用命令上传

python D:/Python/Python36/Scripts/scrapyd-deploy **localhost** -p **imagesDown**

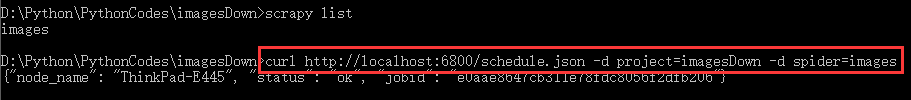


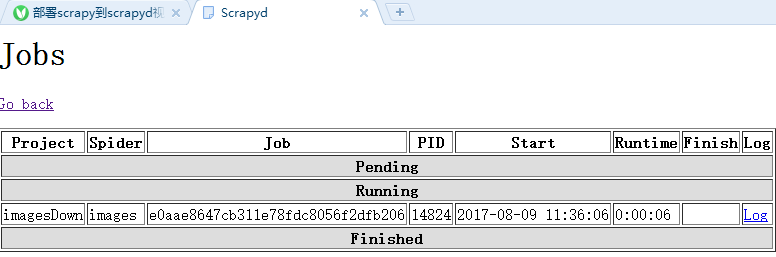




## 对spider调度运行

**curl http://localhost:6800/schedule.json -d project=imagesDown -d spider=images**

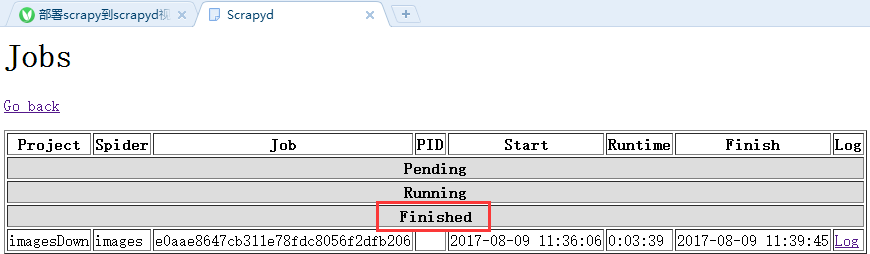




取消spider的执行：

curl http://localhost:6800/**cancel.json** -d project=imagesDown -d job=e0aae8647cb311e78fdc8056f2dfb206





## 列出已经部署的所有的工程

在任意目录下，执行命令：

**curl** [**http://localhost:6800/listprojects.json**](http://localhost:6800/listprojects.json)

**projects 以列表的形式发送回来。**



# 总结

