

## 环境：

Python3.6

Windows 10 专业版 两台

10.100.92.16、10.100.92.42

## Scrapyd 篇

软件介绍：

Scrapyd 是部署和运行 Scrapy 爬虫的应用程序。你可以使用它的 JSON API 部署、控制你的项目，同时也可以对爬虫起到监控作用(但是目前发现主要监控功能为查看爬虫运行的排队、运行和结束状态，对于被反爬、软件报错不能起到监控作用)。

1、如果使用了 virtualenv 虚拟环境，请先进入虚拟环境（我没有使用，所以以下测试均在 python 的默认环境）

```
C:\Users\Administrator>workon spiderkeeper-test
(spiderkeeper-test) C:\Users\Administrator>
```

2、安装 scrapyd（要先安装 scrapy，具体方法请百度）

```
C:\Users\Administrator>pip install scrapyd
...
Successful installed scrapyd-1.2.0
```

3、启动 scrapyd

```
C:\Users\Administrator>scrapyd
2018-04-02T18:22:35+0800 [-] Loading d:\programdata\anaconda3\lib\site-packages\scrapyd\txapp.py...
2018-04-02T18:22:36+0800 [-] Scrapyd web console available at http://0.0.0.0:6800/
2018-04-02T18:22:36+0800 [-] Loaded.
2018-04-02T18:22:36+0800 [twisted.application.app.AppLogger#info]          twistd          17.9.0
(d:\programdata\anaconda3\python.exe 3.6.3) starting up.
2018-04-02T18:22:36+0800 [twisted.application.app.AppLogger#info]          reactor          class:
twisted.internet.selectreactor.SelectReactor.
2018-04-02T18:22:36+0800 [-] Site starting on 6800
2018-04-02T18:22:36+0800 [twisted.web.server.Site#info] Starting factory <twisted.web.server.Site object at
0x00000233F94442E8>
2018-04-02T18:22:36+0800 [Launcher] Scrapyd 1.2.0 started: max_proc=16, runner='scrapyd.runner'
```

#### 4、关闭 scrapyd 命令窗口，配置 scrapyd

找到自己安装 python 的目录，进入（我使用的是 anaconda 集成的 python）D:\ProgramData\Anaconda3\Lib\site-packages\scrapy。

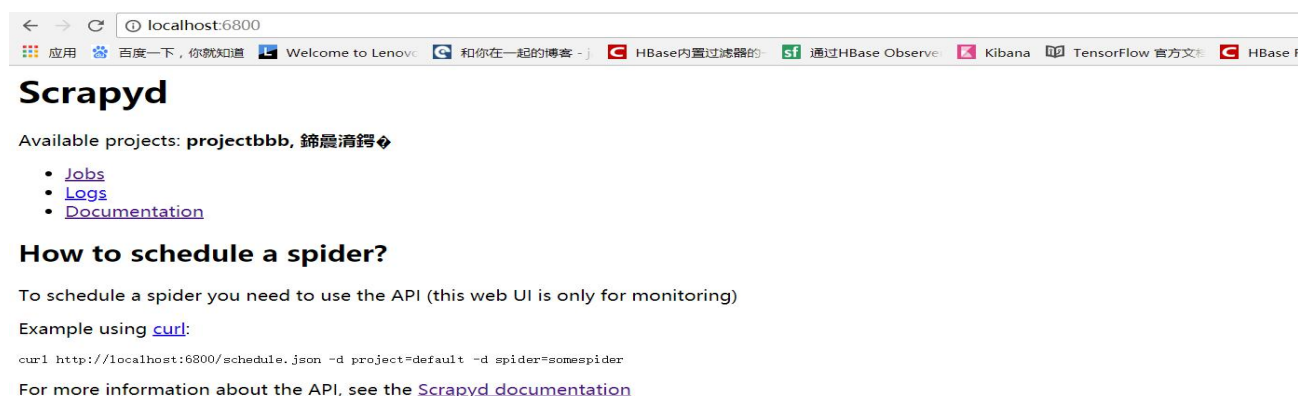
找到目录下的 default\_scrapy.conf 文件，编辑文件，将其中的 bind\_address 127.0.0.1 修改为 0.0.0.0，即可以让所有机器连接该机器的该软件的端口。当然你也可以根据需要修改端口号。（该部分主要为分布式做准备，如果是单机使用也可以不用修改）

另，如果服务器上运行了防火墙，请记得将运行的端口号加入到防火墙例外规则中。

#### 5、再次运行

```
C:\Users\Administrator>scrapyd
2018-04-02T18:22:35+0800 [-] Loading d:\programdata\anaconda3\lib\site-packages\scrapy\txapp.py...
2018-04-02T18:22:36+0800 [-] Scrapyd web console available at http://0.0.0.0:6800/
2018-04-02T18:22:36+0800 [-] Loaded.
2018-04-02T18:22:36+0800 [twisted.application.app.AppLogger#info] twisted 17.9.0
(d:\programdata\anaconda3\python.exe 3.6.3) starting up.
2018-04-02T18:22:36+0800 [twisted.application.app.AppLogger#info] reactor class:
twisted.internet.selectreactor.SelectReactor.
2018-04-02T18:22:36+0800 [-] Site starting on 6800
2018-04-02T18:22:36+0800 [twisted.web.server.Site#info] Starting factory <twisted.web.server.Site object at
0x00000233F94442E8>
2018-04-02T18:22:36+0800 [Launcher] Scrapyd 1.2.0 started: max_proc=16, runner='scrapyd.runner'
```

用浏览器进入 web ui:



← → ↻ localhost:6800

应用 百度一下，你就知道 Welcome to Lenovo 和你在一起的博客 HBase内置过滤器的 sf 通过HBase Observe Kibana TensorFlow 官方文档 HBase R

## Scrapyd

Available projects: **projectbbb**, 锦晨清鐸

- [Jobs](#)
- [Logs](#)
- [Documentation](#)

### How to schedule a spider?

To schedule a spider you need to use the API (this web UI is only for monitoring)

Example using [curl](#):

```
curl http://localhost:6800/schedule.json -d project=default -d spider=somespider
```

For more information about the API, see the [Scrapyd documentation](#)

#### 6、其他

Scrapyd 同时也提供了丰富的 JSON API，你可以通过 curl 对接口进行操作。如果系统没有集成 curl，请百度具体安装方法。

例子：

查看所有项目(在一个新的控制台窗口)：

```
C:\Users\Administrator>curl http://localhost:6800/listprojects.json
```

返回值:

```
{"status": "ok", "projects": ["myproject", "otherproject"]}
```

查看更多 demo 请访问:<http://scrapyd.readthedocs.io/en/stable/api.html> 进行查看。

同时, 如要部署到服务器运行爬虫, 你还需要将你的 scrapy 爬虫的相关依赖、第三方库进行安装。

## Spiderkeeper 篇

软件介绍:

SpiderKeeper 是一款管理爬虫的软件, 和 [scrapinghub](#) 的部署功能差不多, 能多台服务器部署爬虫, 定时执行爬虫, 查看爬虫日志, 查看爬虫执行情况等功能。该软件基于 scrapyd 的 JSON API 进行定制化开发, 相比于 scrapyd 简陋的 web ui, 该软件更加美观、友好, 可以对多个 scrapyd 里的多个项目进行管理。

项目源码地址: <https://github.com/DormyMo/SpiderKeeper>

本攻略中着重讲的是分布式部署。

### 1、安装 spiderkeeper

首先, 你需要在每台服务器上按照 scrapy 篇 的 1 到 5 流程进行安装。

其次, 找一台你觉得 ip 长的好看的主机作为 master。

最后, 进入正题, 在 master 主机(该主机可同时安装 scrapyd, 不要浪费资源哦~)安装 spiderkeeper。

```
C:\Users\Administrator>pip install spiderkeeper
```

```
...
```

```
Successful installed .....
```

### 2、运行 spiderkeeper:

```
C:\Users\Administrator>spiderkeeper --server=http://10.100.92.42:6800 --server=http://10.100.92.16:6800
d:\programdata\anaconda3\lib\site-packages\flask_restful\swagger.py:14:
```

```
ExtDeprecationWarning: Importing flask.ext.restful is deprecated, use flask_restful instead.
```

```
    from flask.ext.restful import Resource, fields
```

```
-----
INFO in run [d:\programdata\anaconda3\lib\site-packages\SpiderKeeper\run.py:22]:
```

```
SpiderKeeper  startd    on    0.0.0.0:5000    username:admin/password:admin    with    scrapyd
servers:http://10.100.92.42:6800,http://10.100.92.16:6800
```

```
-----
2018-04-03  11:15:45,505  -  SpiderKeeper.app  -  INFO  -  SpiderKeeper  startd  on  0.0.0.0:5000
username:admin/password:admin with scrapyd servers:http://10.100.92.42:6800,http://10.100.92.16:6800
```

(注意: 启动时也可以加入 `--port=89791` 来定制端口, `--server` 可以添加一个, 也可以添加多台服务器上 scrapyd 的 server, 在 `http://10.100.92.42:6800` 的最后不加 `'`, 不然会导致软件无法正常使用。启动时你可以在一个指定的文件夹开启一个新控制台来运行, 因为软件会在默认路径创建一个 sqlite 的 DB 文件以及日

```
C:\Users\Administrator>curl http://localhost:6800/listprojects.json
```

志等文件)

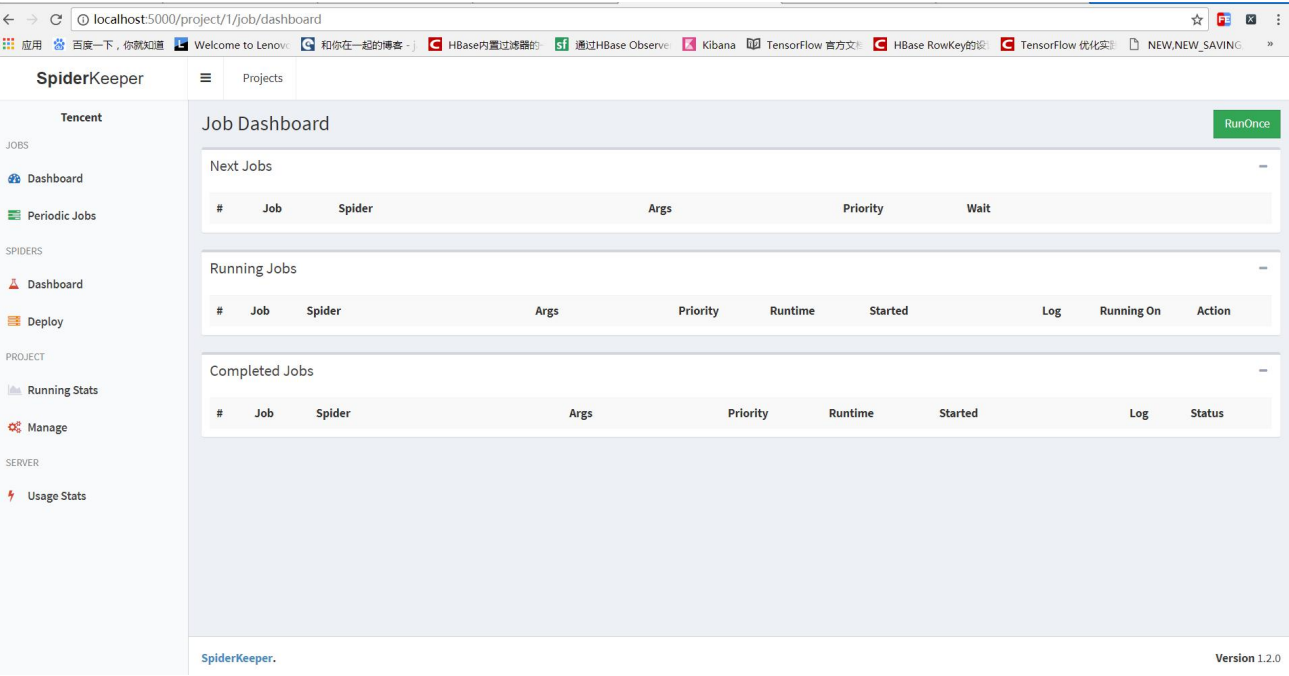
查看更多参数，可以在一个新的控制台输入: spiderkeeper --help 查看。

```
C:\Users\Administrator>spiderkeeper --help
d:\programdata\anaconda3\lib\site-packages\flask_restful_swagger\swagger.py:14:
ExtDeprecationWarning: Importing flask.ext.restful is deprecated, use flask_restful instead.
  from flask.ext.restful import Resource, fields
Usage: spiderkeeper [options]

Admin ui for spider service

Options:
  -h, --help            show this help message and exit
  --host=HOST            host, default:0.0.0.0
  --port=PORT            port, default:5000
  --username=USERNAME    basic auth username ,default: admin
  --password=PASSWORD    basic auth password ,default: admin
  --type=SERVER_TYPE     access spider server type, default: scrapyd
  --server=SERVERS        servers, default: ['http://localhost:6800']
  --database-url=DATABASE_URL
                        SpiderKeeper metadata database default:
                        sqlite:///C:\Users\Administrator\SpiderKeeper.db
  --no-auth              disable basic auth
  -v, --verbose          log level
```

进入 web ui 检测是否启动成功:



### 3、部署项目

① 首先需要在你本机上(你写项目的机器上)安装 scrapy-client

```
C:\Users\Administrator>pip install scrapyd-client
...
Successful installed scrapyd-client-1.1.0
```

注:windows 系统, 在 c:\python36\Scripts(根据你的安装路径来, 下同)下生成的是 scrapyd-deploy, 无法直接在命令行里运行 scrapd-deploy。

需要在 c:\python36\Scripts 下新建一个 scrapyd-deploy.bat, 文件内容如下:

```
@echo off
C:\Python36\python
C:\Python36\Scripts\scrapyd-deploy %*
```

添加环境变量: C:\Python36\Scripts

当然, 你也可以选择不创建.bat 文件, 直接在执行命令时在前边加上 python, 如:

```
C:\Users\Administrator>python scrapyd-deploy --build-egg aa.egg
```

② 生成 egg 文件

第一步, 进入 scrapy 项目根目录, 找到 scrapy.cfg

将 url 前的#去掉, 最后切记不要加'/'号, 有'/'的话要去掉

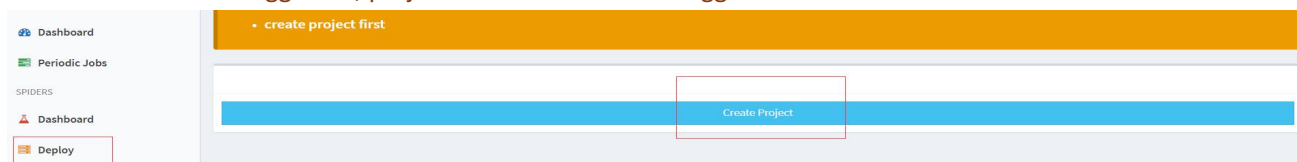
```
[deploy]
url = http://localhost:6800
project = louspider
```

第二步, 在爬虫项目的根目录执行:

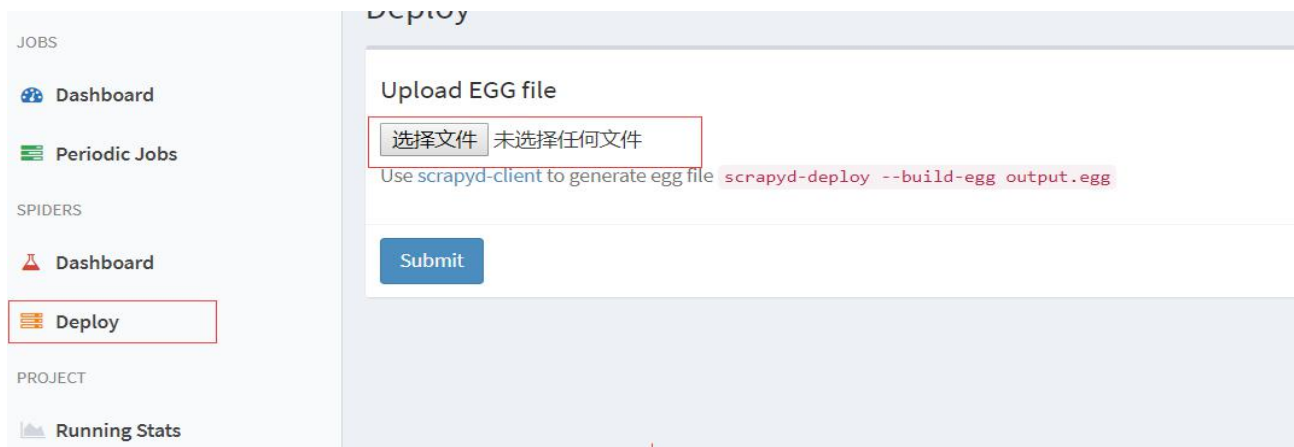
```
F:\Bigdata10veuDuoduo\Python\pyWorkspace\louspider>scrapyd-deploy --build-egg louspider.egg
```

注: egg 文件名尽量保持与项目名相同, 在测试时有因为文件名不相同导致不能使用

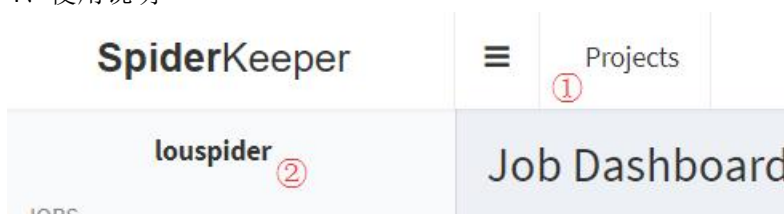
③ 创建项目并上传 egg 文件, project 名称尽量与项目和 egg 文件名称相同



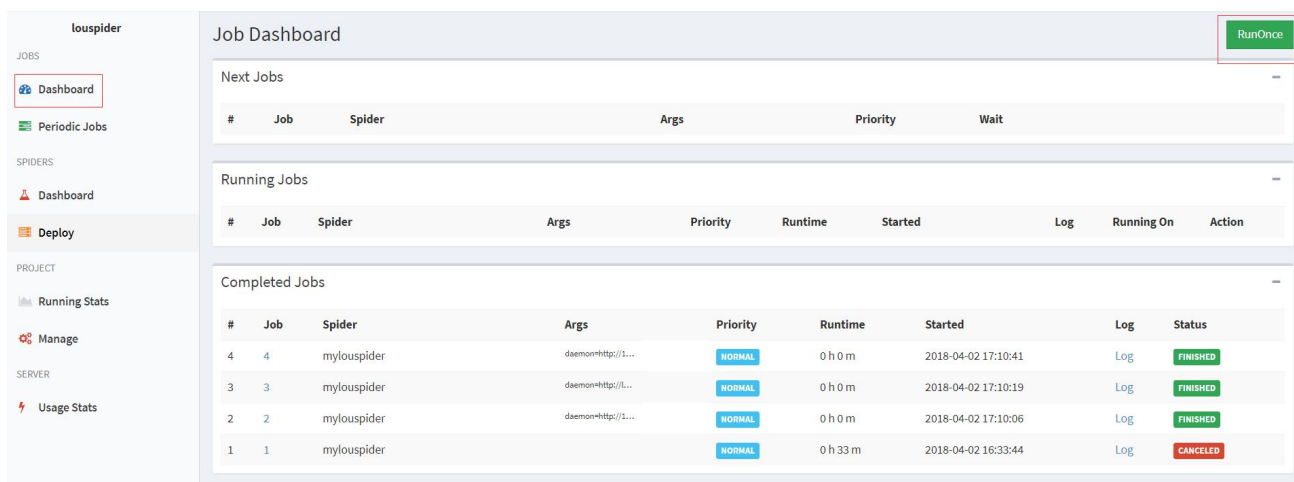
egg 文件上传后会自动部署到每台服务器上, 这个目前暂时没法选择性的部署



#### 4、使用说明



顶部栏①可以选择项目，②是当前的项目名称



Jobs 的 dashboard 部分可以查看当前的 job 和状态，创建一次性的任务可以选择右上角 RunOnce

Run Spider Once

Spider

mylouspider

Priority

Normal

Args

foo=1,bar=2

Chose Daemon

auto

auto

http://localhost:6800

http://10.100.92.42:6800

Close

Create

Spider 是选择要执行的爬虫，地步的 chose daemon 是选择需要运行的机器，其他的可以默认也可以根据需要进行修改。点击 create 就可以执行爬虫了

louspider

JOBS

Dashboard

Periodic Jobs

SPIDERS

Dashboard

Deploy

PROJECT

Running Stats

Manage

SERVER

Usage Stats

Periodic jobs

Add Job

Periodic jobs (Spiders)

#	Month	Day of Month	Day of Week	Hour	Minutes	Spider	Priority	Args	Tags	Enabled	Action
---	-------	--------------	-------------	------	---------	--------	----------	------	------	---------	--------

Jobs 的 periodic jobs 是定时任务版块，点击右上角 add job 可以添加新的 job

Add Periodic Job

Spider

mylouspider

Choose Month

Every Month

Priority

Normal

Choose Day of Week

Every day

Args

foo=1,bar=2

Choose Day of Month

Every day

Choose Hour

Every Hour

Choose Minutes

0

Advanced Options

Chose Daemon

auto

Cron Expressions (m h dom mon dow)

0 \* \* \* \*

Close

Create

这里的设置更 **once** 基本相同，不过可以设置任务执行的周期，右下角也可以写 **cron** 表达式

louspider

JOB

Dashboard

Periodic Jobs

SPIDERS

Dashboard

Deploy

PROJECT

Running Stats

Manage

SERVER

Usage Stats

Spider

Periodic jobs (Spiders)

id	Spider Name	Last Runtime	Avg Runtime
1	mylouspider	2018-04-02 08:33:43	-

这几个版块就不一一介绍了，功能相对容易理解和使用。

该攻略部分摘自网络和 [jiangzl5](#) 的笔记，感谢 [jiangzl5](#)~