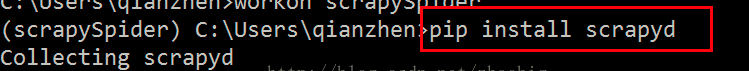
# scrapyd部署爬虫项目

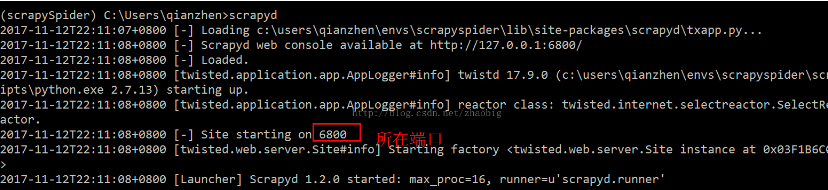
scrapyd是运行scrapy爬虫的服务程序，它支持以http命令方式发布、删除、启动、停止爬虫程序。而且scrapyd可以同时管理多个爬虫，每个爬虫还可以有多个版本。

此部署项目的前提：操作系统-windows、安装有Anaconda、项目python版本-2.7.

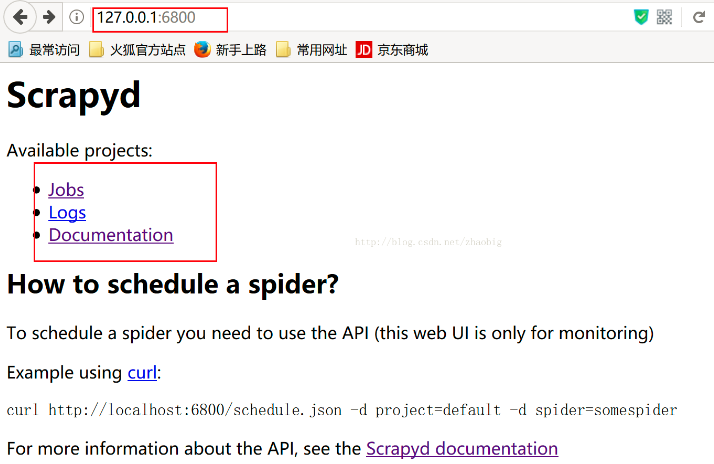
1. 打开Aconada Prompt，新建虚拟环境（方便管理），在虚拟环境中安装scrapy项目需要使用到的包。除了下面提到的包之外，还需要自己在虚拟环境中安装要部署项目中使用的其他包：例如pymongo（3.6.1）、parse（1.8.2）、bs4（0.0.1）、setuptools（39.0.1）、beautifulsoup（4.6.0）、pywin32（222）、requests（2.18.1）、scrapy（1.0.7）等等。
2. 在Aconada Prompt中新建虚拟环境的方法及操作虚拟环境所使用的命令请参考：<https://blog.csdn.net/lyy14011305/article/details/59500819>。
3. 在虚拟环境命令框输入activate XXX（所创建虚拟环境名称），接着安装scrapyd模块，scrapyd模块是专门用于部署scrapy项目的，可以部署和管理scrapy项目。执行pip install scrapyd==1.1.1。



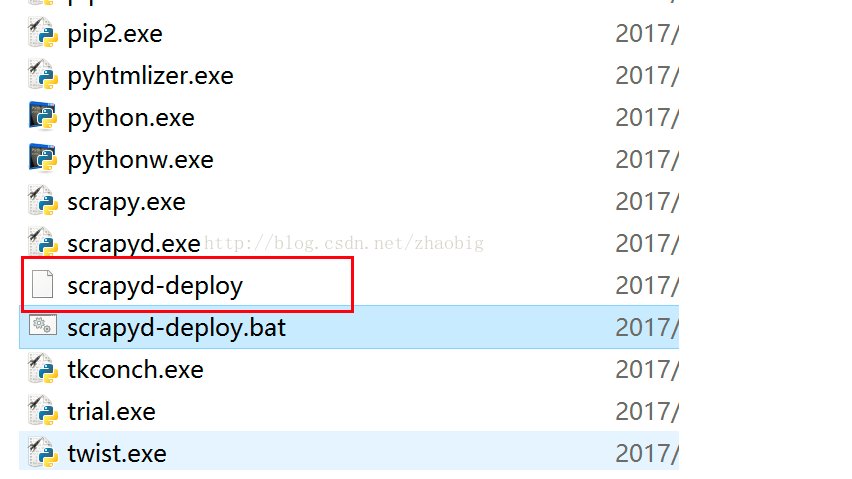
等待安装完成。然后在命令行中输入scrapyd启动scrapyd服务。



在浏览器地址栏中输入：127.0.0.1:6800。显示如下则安装成功。



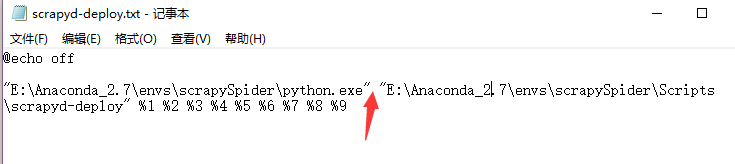
1. 重新打开新的Aconada Prompt进入虚拟环境中，安装scrapy-client。scrapyd-client模块是专门打包scrapy爬虫项目到scrapyd服务中的，进入虚拟环境，执行命令pip install scrapyd-client==1.1.0，安装完成后，在虚拟环境的scripts中会出现scrapyd-deploy无后缀文件，这个scrapyd-deploy无后缀文件是启动文件，在Linux系统下可以远行，在windows下是不能运行的，所以我们需要编辑一下使其在windows可以运行。



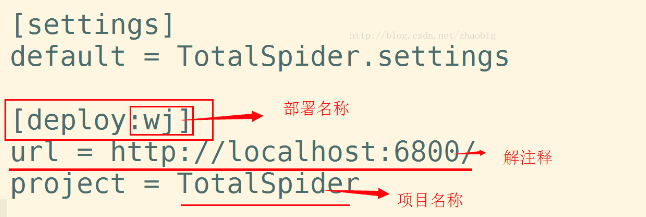
在含有scrapyd-deploy的文件夹中新建一个scrapyd-deploy.bat文件，右键选择编辑，输入以下配置，注意：两个路径之间是空格。

@echo off

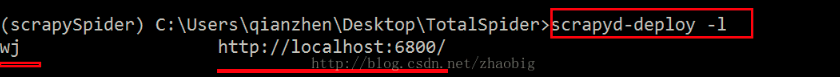
"E:\Anaconda\_2.7\envs\scrapySpider\python.exe""E:\Anaconda\_2.7\envs\scrapySpider\Scripts\scrapyd-deploy" %1 %2 %3 %4 %5 %6 %7 %8 %9



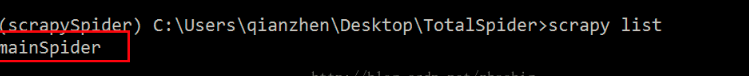
1. 进入虚拟环境，进入到你的爬虫项目中，进入带有scrapy.cfg文件的目录，执行命令scrapyd-deploy，测试scrapyd-deploy是否可以运行，如果出现以下则正常。
2. 打开爬虫项目中的scrapy.cfg文件，这个文件就是给scrapyd-deploy使用的，将url这行代码解掉注释，并且给设置你的部署名称。



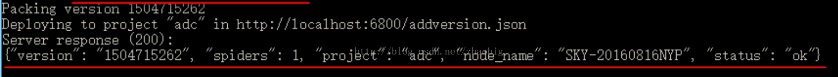
1. 再次执行scrapyd-deploy -l 启动服务，可以看到设置的名称则正确。



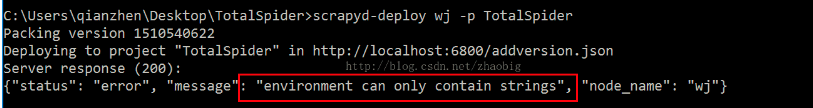
1. 开始打包前，执行一个命令：scrapy list，这个命令执行成功说明可以打包了，如果没执行成功说明还有工作没完成。执行 scrapy list命令，返回了爬虫名称说明一切ok了，如下所示：



1. 到此我们就可以开始打包scrapy项目到scrapyd了，用命令结合scrapy项目中的scrapy.cfg文件设置来打包。执行打包命令： scrapyd-deploy 部署名称 -p 项目名称。如：scrapyd-deploy wj -p TotalSpider。如下显示表示scrapy项目打包成功。



如果出现：

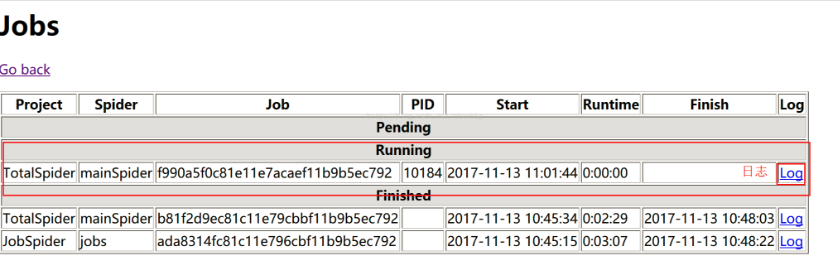


则重新多次提交，还是不可以的话重新启动scrapyd服务或者计算机。

1. 然后在命令行工具（不用进入虚拟环境中）输入开启爬虫命令：curl http://localhost:6800/schedule.json -dproject=项目名称 -d spider=爬虫名称。如果出现如下所示可能会成功。



此时去网页中127.0.0.1:6800查看爬虫运行状态，如果如下所示则一切正常。



以上所有都正常唯独在浏览器中没有正常显示出来，那么查看scrapyd服务中的信息，一般会有报错信息，如果出现：scrapyd spawnProcess not available since pywin32 is not installed，认真查看自己是否已安装pywin32或者安装的是最新版的（卸载安装前期版本稳定性会好一点）。

停止爬虫：curl http://localhost:6800/cancel.json -dproject=scrapy项目名称 -d job=运行ID。例如：curl http://localhost:6800/cancel.json -d project=ArticleSpider -d job=2a9b218a13e011e888cb28d2449bc99e。

1. 每次执行一些列更改的措施后，记得重新启动scrapyd服务器，否则可能会没有任何反应或者继续失败。
2. 参考一下网站：

<https://blog.csdn.net/lyy14011305/article/details/59500819>

<https://blog.csdn.net/u010983763/article/details/77247754>

<https://blog.csdn.net/u013708440/article/details/53425655>

<https://blog.csdn.net/u010983763/article/details/77247754>

<https://blog.csdn.net/zhaobig/article/details/78670176>

<https://blog.csdn.net/x631617479/article/details/72154677>

<https://blog.csdn.net/xiaoquantouer/article/details/53164306>

<https://blog.csdn.net/qq_23079443/article/details/79644476>

Activate scrapyspider :开启虚拟环境

Scrapyd：开启scrappyd服务器（<http://127.0.0.1:6800/>浏览器验证）

Cd F:\SoftProgram\pycharm\JobInformation：进入项目目录（F:）

scrapyd-deploy -l：开启服务

scrapy list：打包前

scrapyd-deploy inf\_spider -p JobInformation：开始打包

执行爬虫：

curl http://localhost:6800/schedule.json -d project= JobInformation -d spider=Boss

curl http://localhost:6800/schedule.json -d project= JobInformation -d spider=Highpin

curl http://localhost:6800/schedule.json -d project= JobInformation -d spider=ZhiLian

停止爬虫：

curl http://localhost:6800/cancel.json -d project= JobInformation -d job=