链滴

docker 部署 scrapy 爬虫（一）

作者：fjun

原文链接：https://ld246.com/article/1531456819128

来源网站：链滴

1. 需求背景

环境搭建是一个繁琐的过程，例如python2.7和python3.6的版本问题, python中一些安装比较麻烦模块如cv2、PIL、pyltp等。我们在本地开发好了一个爬虫项目，需要部署到多台服务器上跑起来，须得在各台服务器上重新搭建爬虫运行环境。

如何解决本地开发完成，马上就能上线运行呢？virtualenv是一种方式，但不适合线上部署运维； 另种更为便捷的方式就是使用docker。docker可以提供操作系统级别的虚拟环境，我们可以将项目制成docker镜像，只要其他机器安装了docker，下载镜像运行即可。docker运行采用虚拟环境，和宿机完全隔离，不必担心环境配置问题或版本冲突问题。

2. 爬虫项目

使用已经完成的一个scrapy爬虫项目，项目地址为：https://github.com/hexiaosong/cnblogs。克项目至本地，项目单独运行:

```
$ workon python2.7
$ cd cnblogs
$ pip install -r requirements.txt
$ export PYTHONPATH=$(pwd)
$ python main.py
```

3. 创建Dockerfile

确保本地已经安装了Docker并正常运行。在项目的根目录下新建一个Dockerfile，文件不加任何后缀，添加以下内容:

```
FROM hexiaosong/python2.7:latest
RUN yum install -y python-devel  gcc gcc-c++ autoconf automake libtool make
ENV PATH /usr/local/bin:$PATH
ADD . /code
VOLUME /data
WORKDIR /code
RUN pip install -i  https://pypi.doubanio.com/simple/  --trusted-host pypi.doubanio.com -r r
```

```
quirements.txt
CMD ["python", "main.py"]
```

● 第一行FROM代表使用的Docker基础镜像，这里我们使用lmurawsk/python2.7的作为基础镜像，此基础上运行爬虫项目

● 第二行RUN是安装python的环境依赖

● 第三行ENV是环境变量设计，将/usr/local/bin:$PATH赋值给PATH，即添加/usr/local/bin环境变路径

● 第四行是添加当前代码到容器的/code目录

● 第五行VOLUME是在容器中创建数据卷，作为数据挂载点，将爬取的数据从容器中取出至本地，后会用到

● 第六行是将/code设置为工作目录

● 第七行CMD是容器启动命令，运行项目

4. 构建镜像

在根目录下运行一下命令:

```
$ docker built -t cnblogs:latest .
```

注意后面的小点

执行过程输出如下所示:

这样数据就说明镜像构建成功，查看已经构建的镜像

```
docker images
```

5. 运行

测试镜像在本地运行,执行一下命令:

```
docker run -v /Users/apple/Downloads/data:/data cnblogs
```

/Users/apple/Downloads为本地存放爬虫结果的路径

这样我们就用镜像新建并运行了一个Docker容器，运行效果和直接跑scrapy项目完全一样, 如下图所：

当运行完成时，可以看到本地目录/Users/apple/Downloads下生成了cnblog.json, 即为爬取结果.

6.拷贝此镜像至服务器

将镜像拷贝成.tar文件

```
docker save -o cnblogs.tar cnblogs:latest
```

将cnblogs.tar拷贝至服务器，确保服务器安装并运行了Docker。载入cnblogs.tar。

```
docker load --input cnblogs.tar cnblogs:latest
```

在服务器上运行镜像

```
docker run -v /home:/data cnblogs
```