Scrapy Workshop

林帅

2014.11.15

Scrapy Workshop

- 介绍 scrapy 和 scrapy cloud
- 写一个简单的爬虫，在本地运行
- 将爬虫布署到 scrapy cloud 上运行

准备

- 操作系统： Linux / MacOS
- 基本的 python 编程知识

Scrapy

- 网络爬虫 101
- 框架

Information

- 采集
- 解析
- 存储
- 访问

爬虫应用

- 搜索引擎
- 比价网站
- 媒体检测

网络爬虫

- 获取网页
- 解析网页
- 其他细节，比如 login, cookie, 速度控制

Scrapy Cloud

- 专门运行爬虫的 PaaS
- 布署: scrapy deploy
- webapi

解析网页内容

- HTML
- CSS selector
- XPath

动手

- 完成一个简单的爬虫
- 目标1：豆瓣电影 http://movie.douban.com
- 目标2：抓取 Pycon China 2014 演讲嘉宾

- sudo pip install scrapy
- 如果没有安装 pip, 先安装 pip
- ubuntu: sudo apt-get install python-pip
- MacOS: sudo port install py27-pip

目标：豆瓣电影

http://movie.douban.com

获取代码

- git clone git://gitcafe.com/lins05/pycon-spiders.git

本地运行

- scrapy crawl movie -o movie.json

完善爬虫代码

- 目前的爬虫只抓取了电影名字，需要动手改代码，抓取更多信息

在 scrapy cloud 上运行

- 注册账号 https://dash.scrapinghub.com
- 创建 project
- 获取 api key

修改 scrapy.cfg

```
[deploy]
url = http://dash.scrapinghub.com/api/scrapyd/
project = 6650
version = GIT
```

上传代码到 scrapy cloud

- scrapy deploy

启动 spider

- https://dash.scrapinghub.com

目标二：pycon speakers

http://cn.pycon.org

创建一个新项目

- scrapy startproject pyconspeakers
- scrapy genspider speakers pycon.org