

爬虫工程师（一）

 DataCastle
数据科学学习社区

造数

联合制作

造数



导师介绍



霹雳黄（ Billy ）

黄震昕，「造数科技」创始人，其提供的云端爬取技术使得更大范围内的数据采集成为可能。2016年获得「明势资本」和「仟跃大数据」的天使轮投资。



本节知识点概述

什么是爬虫

网址

网址的构成
输入网址后发生了什么
翻页后URL的变化

网页源代码

如何查看
包含什么信息
拓展

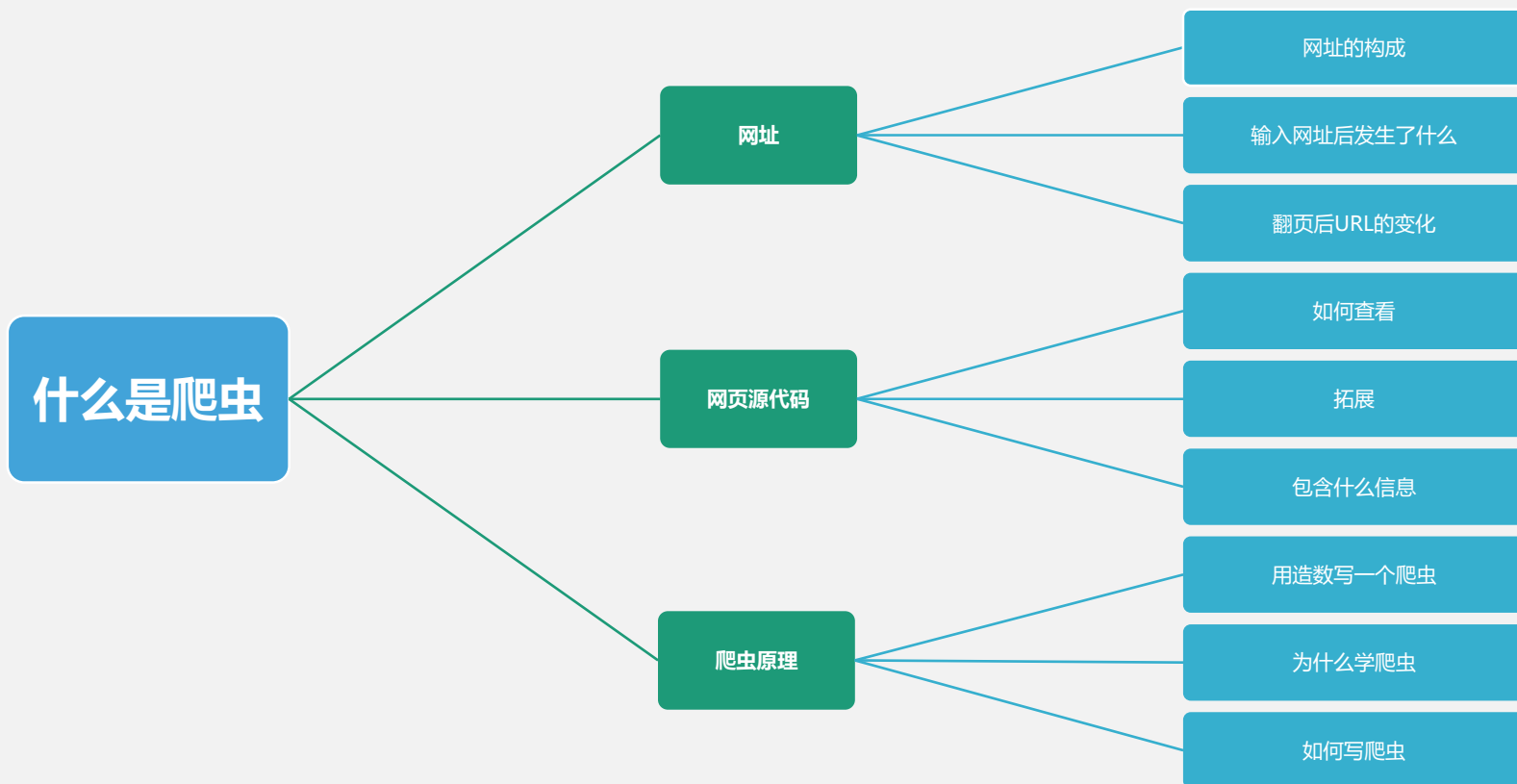
爬虫原理

用造数写一个爬虫
如何写爬虫
为什么学爬虫





本节知识点概述





认识网址的构成

输入网址后发生了什么

http://www.itjuzi.com/company





URL如何记录翻页信息

翻页后URL的变化

第一页：<http://www.itjuzi.com/company>

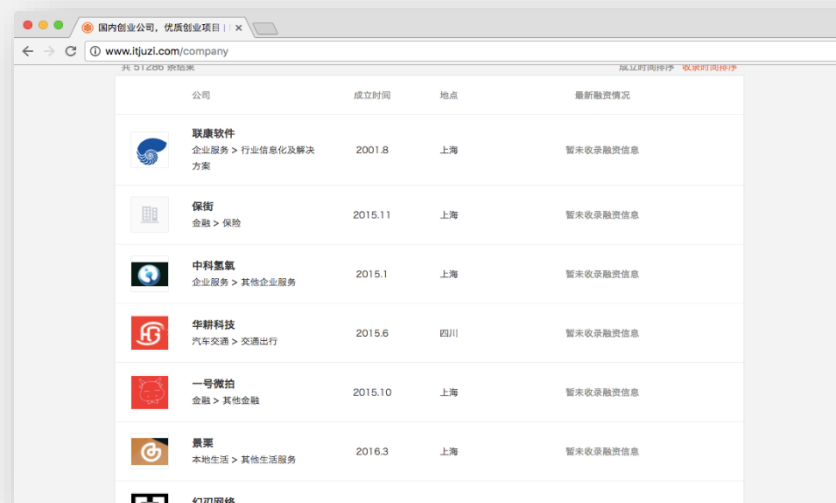
第二页：<http://www.itjuzi.com/company?page=2>








第三页：<http://www.itjuzi.com/company?page=3>

.....

第N页：???

URL改变的翻页



公司	成立时间	地点	最新融资情况
 联康软件 企业服务 > 行业信息化及解决方案	2001.8	上海	暂未收录融资信息
 保衡 金融 > 保险	2015.11	上海	暂未收录融资信息
 中科基氧 企业服务 > 其他企业服务	2015.1	上海	暂未收录融资信息
 华研科技 汽车交通 > 交通出行	2015.6	四川	暂未收录融资信息
 一号微拍 金融 > 其他金融	2015.10	上海	暂未收录融资信息
 晨翼 本地生活 > 其他生活服务	2016.3	上海	暂未收录融资信息
 幻刃网络			



URL如何记录翻页信息

URL不改变的翻页

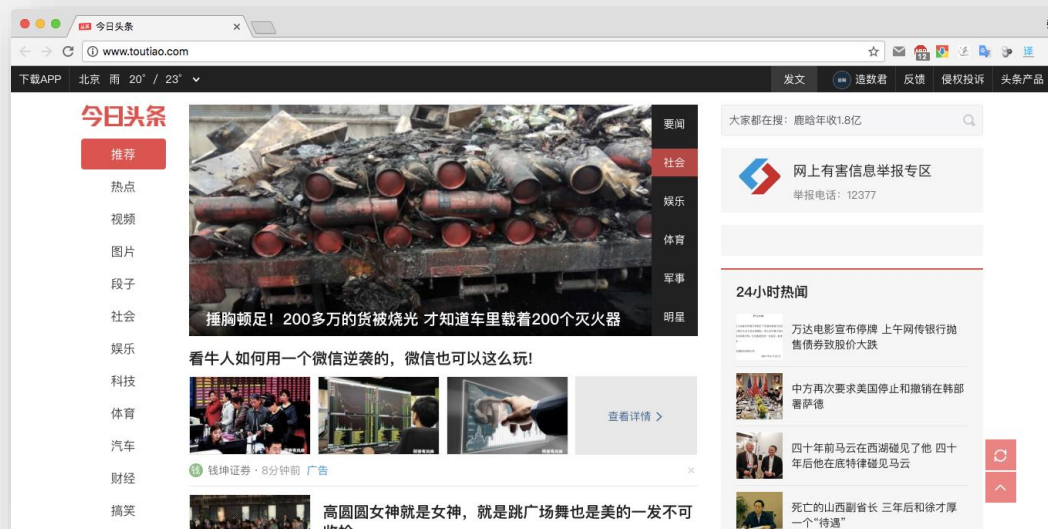
第一页：<http://www.toutiao.com/>

第二页：<http://www.toutiao.com/>

第三页：<http://www.toutiao.com/>

.....

第N页：???





从网页源代码查看信息

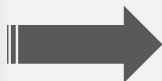
返回
前进
重新加载

存储为...
打印...
投射...
翻成中文 (简体)

印象笔记·剪藏

显示网页源代码

检查



```
view-source:https://www.zaoshu.io

1 <!DOCTYPE html>
2 <html>
3   <head>
4     <meta charset="utf-8" />
5     <meta http-equiv="X-UA-Compatible" content="IE=Edge">
6     <meta name="viewport" content="user-scalable=no, initial-scale=1, maximum-scale=1, minimum-scale=1, target-densitydpi=medium-dpi" />
7     <meta name="Keywords" content="造数,造数科技,爬虫工具,云爬虫,爬虫,采集,网页采集,数据采集,智能云爬虫,网页抓取,网络爬虫">
8     <meta name="description" content="造数,免费数据采集工具,是国内最好用的云爬虫,为个人提供数据采集和开发者功能,为企业提供专业化的数据抓取,数据实时监控服务。联系电话 13366254026">
9     <title>造数 - 新一代智能云爬虫</title>
10    <link rel="stylesheet" type="text/css" href="app.css?v=1498031447.fc3f833fc9253700"/>
11    <link rel="shortcut icon" href="img/favicon.ico">
12    <link rel="icon" type="images/png" sizes="16x16" href="img/Icon_16x16.png">
13    <link rel="icon" type="images/png" sizes="32x32" href="img/Icon_32x32.png">
14    <link rel="icon" type="images/png" sizes="64x64" href="img/Icon_64x64.png">
15    <link rel="icon" type="images/png" sizes="96x96" href="img/Icon_96x96.png">
16    <meta name="baidu-site-verification" content="8Wp57Bg5qd" />
17    <meta name="google-site-verification" content="FAlMS_2q6vxxvKBnS8P0kqv8913oNEPByGWPb1C7o1eQ" />
18    <meta name="360-site-verification" content="5b4884d9919f94881baa887e6deb78ef" />
19    <!--[if lte IE 9]>
20    <script>
21      window.isLteIE9 = true;
22    </script>
23    <link rel="stylesheet" type="text/css" href="ie.css?v=1498031447.fc3f833fc9253700"/>
24    <![endif]-->
25  </head>
26  <body>
27    <div class='main'>
28      <div class='top'>
29        <div class='content'>
30          <div class='logo'>
31            <img src='img/logo.svg?v=1498031447.fc3f833fc9253700' onerror='this.src='img/logo.gif';this.onerror=null;' width='72' height='30' />
32          </div>
33          <div class='signup'>注册</div>
34          <div class='login'>登录</div>
35          <div class='btn white goDashboard'>我的控制面板</div>
36          <div class='menus'>
37            <a class='menuItem' href='/'>首页</a>
38            <a class='menuItem' href='/features.html'>功能</a>
39            <a class='menuItem' href='/pricing.html'>价格</a>
40          </div>
41        </div>
42      </div>
43    </div>
44  </body>
45 </html>
```



DataCastle
数据科学学习社区

造数



网页代码和浏览器展示的信息

网页源代码



浏览器翻译

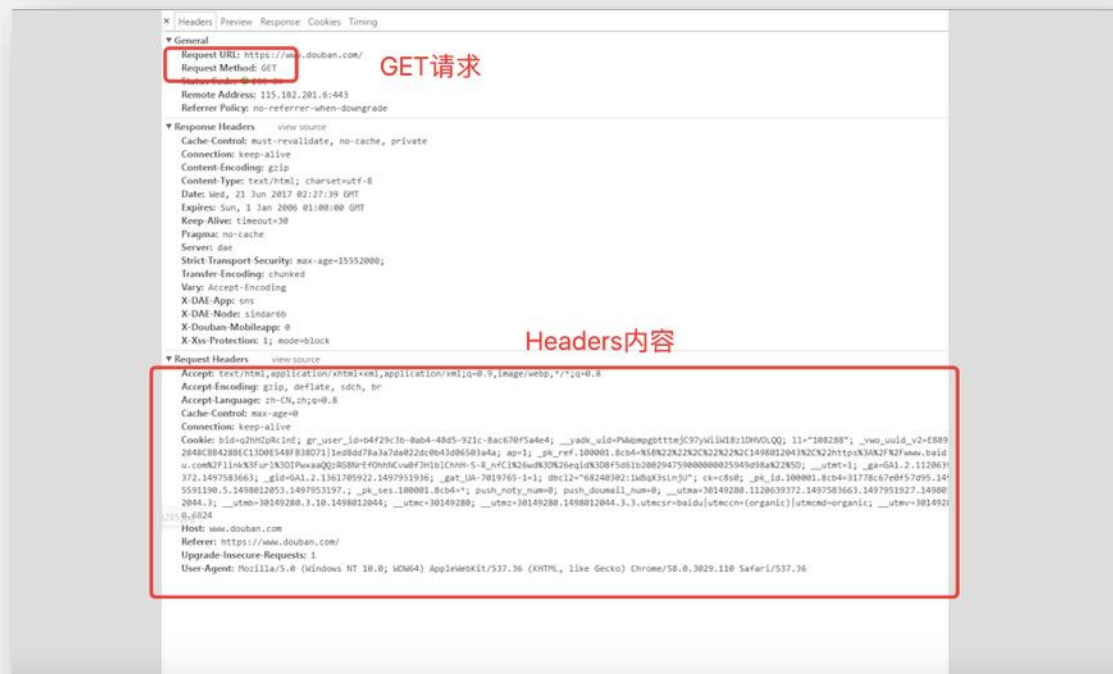
```
<div class='login'>登录</div>
<div class='btn white goDashboard'>我的控制面板</div>
<div class='menus'>
  <a class='menuItem' href='/'>首页</a>
  <a class='menuItem' href='/features.html'>功能</a>
  <a class='menuItem' href='/pricing.html'>价格</a>
  <a class='menuItem' href='/business.html'>商务合作</a>
  <a class='menuItem' href='https://help.zaoshu.io/hc/zh-cn/sections/115002096728-FAQ' target='_blank'>帮助中心</a>
  <a class='menuItem' href='https://forum.zaoshu.io/' target='_blank'>社区</a>
  <a class='menuItem' href='https://zhuanlan.zhihu.com/c_80658853' target='_blank'>博客</a>
</div>
```





如何查看网页的请求

1. 单击鼠标右键，检查（或者F12）
2. 选择Network，刷新页面
3. 选中ALL下面的第一个链接



HTTP 协议

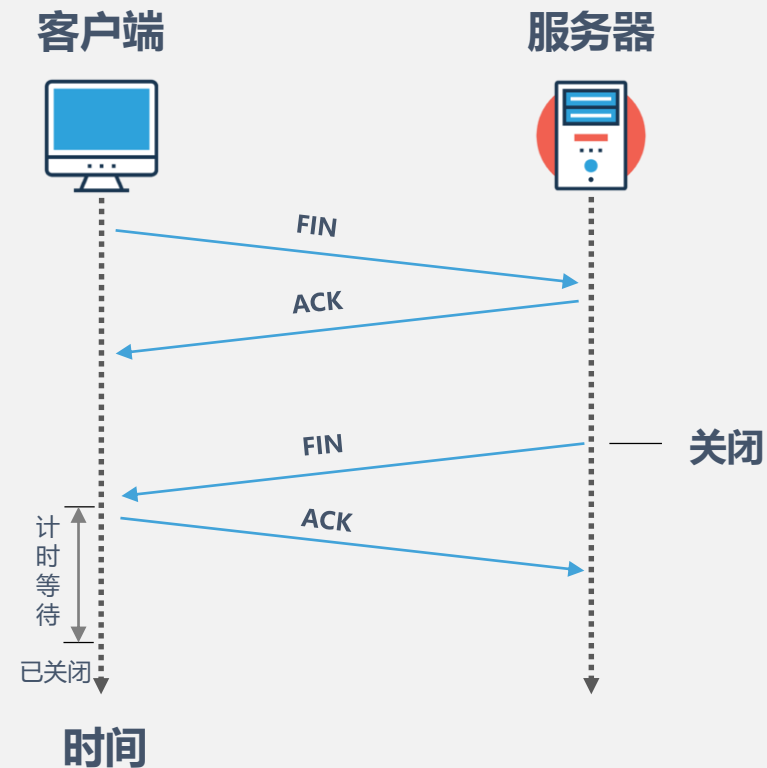
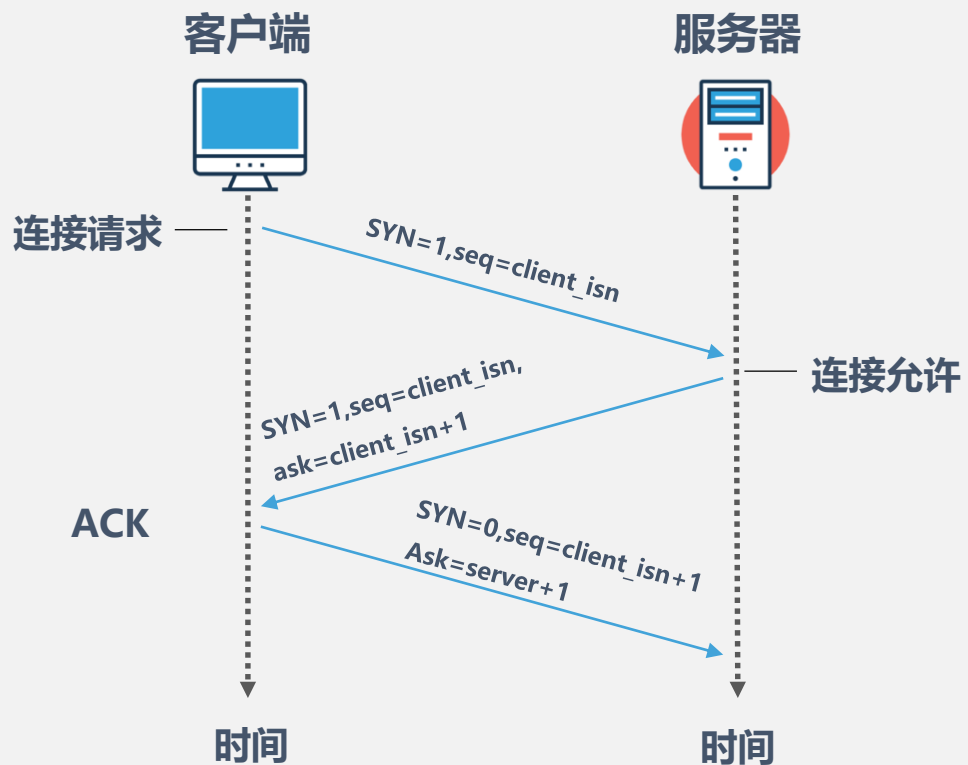
1. 浏览器作为HTTP客户端通过URL向HTTP服务
2. Web服务器根据接收到的请求后，向客户端发送响应信息





TCP 连接的建立与断开

建立TCP需要三次握手才能建立，而断开连接则需要四次握手





爬虫原理

使用爬虫可以获得什么，以造数为案例

<https://www.zaoshu.io/>



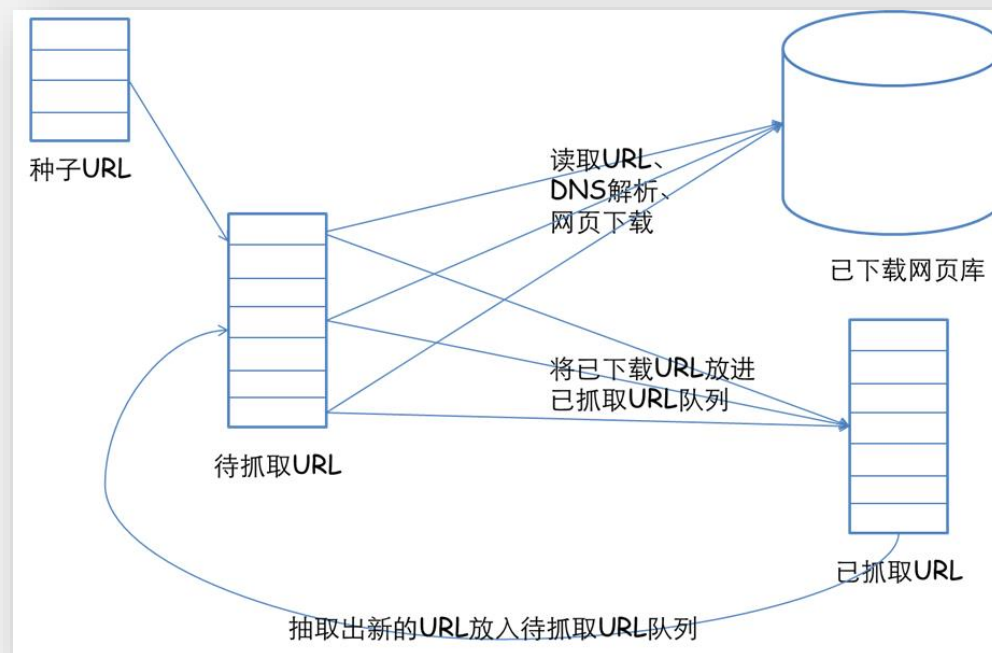
DataCastle
数据科学学习社区

造数



通用的网络爬虫框架

1. 选取挑选的种子URL；
2. 将这些URL放入待抓取URL队列；
3. 取出待抓取的URL，下载，存储进已下载网页库中。此外，将这些URL放进已抓取URL队列。
4. 分析已抓取队列中的URL，并且将URL放入待抓取URL队列，从而进入下一个循环。





如何写爬虫

获得源码

- Urllib、Requests

解析源码

- BeautifulSoup、正则表达式

保存数据

- Xlrd、MySQL、MongoDB



为什么学习爬虫

- 高薪和良好的发展前景
- 获得想要的数​​据
- 有助于搜索引擎优化
- 定制一个搜索引擎

<p>爬虫工程师 [北京 · 亮马桥] 15:41发布</p> <p>15k-30k 经验3-5年 / 本科</p> <p>高级 中级 分布式</p>	<p>国美互联网</p> <p>电子商务 / 上市公司</p> <p>“技术氛围,弹性时间,规模互联网”</p>
<p>爬虫工程师 [成都 · 高新区] 09:47发布</p> <p>18k-30k 经验1-3年 / 本科</p> <p>资深 高级 软件开发 php python</p>	<p>趣玩网</p> <p>移动互联网,电子商务 / 成熟型(D轮及以上)</p> <p>“核心项目,行业前景,福利齐全,弹性工作”</p>
<p>爬虫工程师 [深圳 · 科技园] 1天前发布</p> <p>10k-20k 经验1-3年 / 本科</p> <p>Python 大数据</p>	<p>华海乐盈</p> <p>移动互联网 / 成熟型(不需要融资)</p> <p>“扁平化管理,弹性工作制,绩效奖金,培训”</p>
<p>爬虫工程师 [深圳 · 南头] 10:16发布</p> <p>15k-25k 经验不限 / 本科</p> <p>数据分析 Python 大数据 scrapy Gevent</p>	<p>极光</p> <p>移动互联网,金融 / 成熟型(C轮)</p> <p>“14薪,五险一金,弹性工作制,技术氛围好”</p>

拉勾网：爬虫工程师职位信息



使用造数爬取淘宝

正在爬取: 手机_淘宝搜索
查看使用说明

完成创建

返回

全场12期免息 R11娇兰热力红礼盒

6.30 开售 抢先预约

客服咨询

已选中 6 列数据, 爬取结果预览如下

匹配效果不理想? 告诉我们这个问题

列-1	列-2	列-3	列-3-链接
【旗舰新品】OPPO R11 全网通前后2000万 指纹识别拍照手机r11r9s	3199.00	oppo手机官方旗舰店	https://click.simba.taobao.com/cc_im?...
Xiaomi/小米 红米 手机 4X 全网通4g超... 你智能指纹识别学生手机	699.00	小米官方旗舰店	https://store.taobao.com/shop/view_sh...
【旗舰新品】OPPO R11 全网通前后2000万 指纹识别拍照手机r11r9s	3199.00	oppo手机官方旗舰店	https://store.taobao.com/shop/view_sh...



使用 Python 爬取淘宝

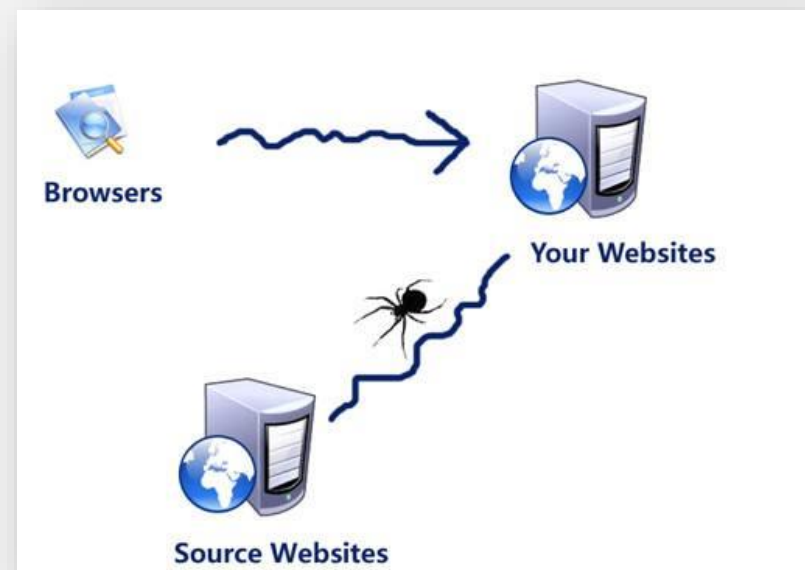
```
1  # -*- coding:utf-8 -*-
2
3  import urllib
4  import urllib2
5  import re
6  import tool
7  import os
8
9  #抓取MM
10 class Spider:
11
12     #页面初始化
13     def __init__(self):
14         self.siteURL = 'http://mm.taobao.com/json/request_top_list.htm'
15         self.tool = tool.Tool()
16
17     #获取索引页面的内容
18     def getPage(self, pageIndex):
19         url = self.siteURL + "?page=" + str(pageIndex)
20         request = urllib2.Request(url)
21         response = urllib2.urlopen(request)
22         return response.read().decode('gbk')
23
24     #获取索引界面所有MM的信息，list格式
25     def getContents(self, pageIndex):
26         page = self.getPage(pageIndex)
27         pattern = re.compile('<div class="list-item" .*?pic-word.*?<a href="(.*?)".*?<img src="(.*?)".*?<a class="lady-'
28         items = re.findall(pattern, page)
```





尝试完成一下作业

- 用自己的话，说说什么是爬虫
- 查看网页的源代码
- 了解网页的翻页规则
- 使用造数爬取某个网站的数据



爬虫工程师

更多数据科学课程，上DC官网：www.pkbigdata.com



关注 DataCastle



关注造数



造数

