

成果形式： 实践报告

成果名称： 关于 python 网络爬虫的实现

一. 实践目的

(一)、为什么学习爬虫

1. 企业生产的用户数据：大型互联网公司海量用户，所以他们积累数据有天然的优势。有数据意识的中小型企业，也开始积累的数据。

2. 数据管理咨询公司：通常这样的公司有很庞大的数据采集团队，一般会通过市场调研、问卷调查、固定的样本检测和各行各业的公司进行合作、专家对话（数据积累很多年了，最后得出科研结果）来采集数据。

3. 政府/机构提供的公开数据：政府通过各地政府统计上报的数据进行合并；机构都是权威的第三方网站。

4、学习如何采集数据：

一般会通过市场调研, 问卷调查, 固定的样本检测和各行各业的公司进行合作, 专家对话（数据积累很多年了，最后得出科研结果）来采集数据。

5、了解社会对计算机新技能的运用：

社会的不断发展导致新技能的强烈需求，原始的计算技能已经不再适用新时代的发展，学习一门新的适应时代潮流的技能有助于我们在社会更好立足。

6、锻炼自主上机操作的能力：

现今很多学生都沉迷于手机和电子游戏中，很少时间来接触和学习新的计算机语言，短时间的实践和学习可以帮助我们提升上机操作能力。

7、解释爬虫

```
<!DOCTYPE html>
```

```
<html lang="zh-CN">
```

```
<head>
```

```

<meta charset="utf-8" />

<meta name="renderer" content="webkit"/>

<meta http-equiv="X-UA-Compatible" content="IE=Edge"/>

<link rel="dns-prefetch" href="//g.alicdn.com"/>

<link rel="dns-prefetch" href="//img.alicdn.com"/>

<link rel="dns-prefetch" href="//gm.mmstat.com"/>

<link rel="dns-prefetch" href="//ald.taobao.com"/>

<link rel="dns-prefetch" href="//bar.tmall.com"/>

<link href="//img.alicdn.com/tfs/TB1X1F3RpXXXXc6XXXXXXXXXXXX-16-16.png"
rel="shortcut icon" type="image/x-icon"/>

<link title="天猫 Tmall.com"
href="//g.alicdn.com/mui/global/1.2.35/file/search.xml"
type="application/opensearchdescription+xml" rel="search"/>

<meta name="spm-id" content="a2156.1676643"/>

<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1" />

<meta name="renderer" content="webkit" />

<title>淘宝网 - 淘! 我喜欢</title>

<base target="_blank"/>

<link rel="stylesheet"
href="//g.alicdn.com/??mui/global/3.0.31/global.css,mui/global/3.0.31/responsive.css,
zebra-pages/portal-4713/1.1.17/pc/index.css" />

<script src="

```

//g.alicdn.com/??kissy/k/1.4.16/seed-min.js,mui/global/3.0.31/global-pc.js,mui/global/3.0.31/global.js,mui/globalmodule/3.0.83/seed.js,mui/btscfg-g/3.0.0/index.js,mui/bucket/3.0.4/index.js,mui/globalmodule/3.0.83/global-mod-pc.js,mui/globalmodule/3.0.83/global-mod.js,mui/global/3.0.31/responsive.js,mui/kissy-polyfill/4.0.16/index.js,zebra-pages/portal-4713/1.1.17/pc/index.js,zebra-pages/portal-4713/1.1.17/seed.js"></script>

<meta name="spm-id" content="a21bo" />

<meta name="description" content="淘宝网 - 亚洲较大的网上交易平台，提供各类服饰、美容、家居、数码、话费/点卡充值…数亿优质商品，同时提供担保交易(先收货后付款)等安全交易保障服务，并由商家提供退货承诺、破损补寄等消费者保障服务，让你安心享受网上购物乐趣！" />

<meta name="aplus-xplug" content="NONE">

<metaname="keyword" content="淘宝, 淘宝, 网上购物, C2C, 在线交易, 交易市场, 网上交易, 交易市场, 网上买, 网上卖, 购物网站, 团购, 网上贸易, 安全购物, 电子商务, 放心买, 供应, 买卖信息, 网店, 一口价, 拍卖, 网上开店, 网络购物, 打折, 免费开店, 网购, 频道, 店铺" />

(二)、爬虫的分类：根据使用场景：分为 通用爬虫 聚焦爬虫

1、通用爬虫

通用网络爬虫是搜索引擎得到系统的重要组成部分。主要目的是将互联网上的网页下载到本地，形成一个互联网内容的镜像备份。

通用搜索引擎（Search Engine）工作原理

通用网络爬虫 从互联网中搜集网页，采集信息，这些网页信息用于为搜索引擎建立索引从而提供支持，它决定着整个引擎系统的内容是否丰富，信息是否即时，因此其性能的优劣直接影响着搜索引擎的效果。

第一步：抓取网页

搜索引擎网络爬虫的基本工作流程如下：

①首先选取一部分的种子 URL，将这些 URL 放入待抓取 URL 队列；

②取出待抓取 URL，解析 DNS 得到主机的 IP，并将 URL 对应的网页下载下来，存储进已下载网页库中，并且将这些 URL 放进已抓取 URL 队列。

③分析已抓取 URL 队列中的 URL，分析其中的其他 URL，并且将 URL 放入待抓取 URL 队列，从而进入下一个循环....

搜索引擎如何获取一个新网站的 URL：

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/326011211001011010>