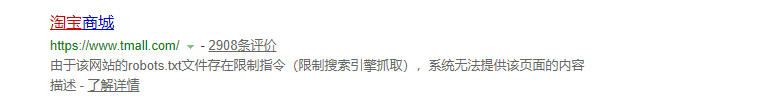
爬虫： 爬取网络的程序，

爬虫的分类，分为通用爬虫和聚焦爬虫。

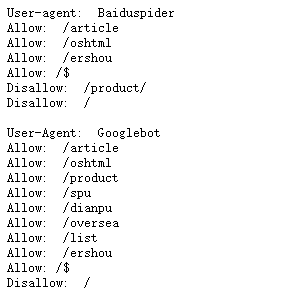
通用爬虫： 搜索引擎的爬虫，robots.txt

Cache ,百度服务器的缓存，

Robots.txt



<https://www.taobao.com/robots.txt>



Robots.txt协议规定通用爬虫，运行爬取网页的作用范围。

聚焦爬虫： 程序员写爬虫，

爬虫界两大阵营：

爬虫 —— 反爬虫 —— 反反爬虫

了解反爬虫， 携程 反爬虫

<https://segmentfault.com/a/1190000005840672?utm_source=tag-newest>

始终胜利的总会是爬虫这一方。

# 案例1 爬取百度贴吧

分析网页：

分析url:

第一页：

[http://tieba.baidu.com/f?kw=%E5%8D%8E%E5%8D%97%E5%86%9C%E4%B8%9A%E5%A4%A7%E5%AD%A6&pn=0](http://tieba.baidu.com/f?kw=%E5%8D%8E%E5%8D%97%E5%86%9C%E4%B8%9A%E5%A4%A7%E5%AD%A6&ie=utf-8&pn=0)

第二页

[http://tieba.baidu.com/f?kw=%E5%8D%8E%E5%8D%97%E5%86%9C%E4%B8%9A%E5%A4%A7%E5%AD%A6&pn=50](http://tieba.baidu.com/f?kw=%E5%8D%8E%E5%8D%97%E5%86%9C%E4%B8%9A%E5%A4%A7%E5%AD%A6&ie=utf-8&pn=50)

第三页：

<http://tieba.baidu.com/f?kw=%E5%8D%8E%E5%8D%97%E5%86%9C%E4%B8%9A%E5%A4%A7%E5%AD%A6&pn=100>

第四页：

<http://tieba.baidu.com/f?kw=%E5%8D%8E%E5%8D%97%E5%86%9C%E4%B8%9A%E5%A4%A7%E5%AD%A6&pn=150>

第一步 爬取源码

|  |
| --- |
| **from** urllib **import** request  **def** loadPage(url):  req = request.Request(url) *#构建请求对象* response = request.urlopen(req) *#发送请求,得到响应对象* html = response.read() *#获取响应对象的内容* **return** html |
| *#定义一个程序的入口* **if** \_\_name\_\_ == **'\_\_main\_\_'**:  url = **'http://tieba.baidu.com/f?kw=%E5%8D%8E%E5%8D%97%E5%86%9C%E4%B8%9A%E5%A4%A7%E5%AD%A6&pn=0'** html = loadPage (url)  **html = html.decode('utf-8')**  print(html) |

第二步，将源码写入到本地

|  |
| --- |
| **from** urllib **import** request  **def** loadPage(url):  req = request.Request(url) *#构建请求对象* response = request.urlopen(req) *#发送请求,得到响应对象* html = response.read() *#获取响应对象的内容* **return** html  **def** writePage(html,filename):  *''' 将爬取回来的内容，写到文本中'''* **with** open(filename,**'w'**,encoding=**'utf-8'**) **as** f:  f.write(html) |
| *#定义一个程序的入口* **if** \_\_name\_\_ == **'\_\_main\_\_'**:  url = **'http://tieba.baidu.com/f?kw=%E5%8D%8E%E5%8D%97%E5%86%9C%E4%B8%9A%E5%A4%A7%E5%AD%A6&pn='**   html = loadPage(url)  html = html.decode(**'utf-8'**) *# 对unicode编码进行解码* print(html)  **filename = '华农吧第1页.html'   writePage(html, filename)** |

第三步，设置翻页功能

|  |
| --- |
| **from** urllib **import** request  **def** loadPage(url):  req = request.Request(url) *#构建请求对象* response = request.urlopen(req) *#发送请求,得到响应对象* html = response.read() *#获取响应对象的内容* **return** html  **def** writePage(html,filename):  *''' 将爬取回来的内容，写到文本中'''* **with** open(filename,**'w'**,encoding=**'utf-8'**) **as** f:  f.write(html)  **def** tiebaSpider(url):  *''' 输入起始页- 终止页 爬取贴吧的页面范围'''* beginPage = int(input(**'请输入起始页：'**))  endPage = int(input(**'请输入结束页：'**))  **for** page **in** range(beginPage,endPage+1):  pn = (page-1)\*50  fullurl = url + str(pn)  print(fullurl)   html = loadPage(fullurl)  html = html.decode(**'utf-8'**) *# 对unicode编码进行解码* print(html)  filename = **'华农吧第%s页.html'**%page  writePage(html, filename) |
| *#定义一个程序的入口* **if** \_\_name\_\_ == **'\_\_main\_\_'**:  url = **'http://tieba.baidu.com/f?kw=%E5%8D%8E%E5%8D%97%E5%86%9C%E4%B8%9A%E5%A4%A7%E5%AD%A6&pn='** tiebaSpider(url) |