

[Python学习] 简单网络爬虫抓取博客文章及思想介绍

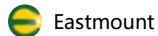
原创 Eastmount 最后发布于2014-10-04 16:33:43 阅读数 13704 ☆ 收藏

编辑 展开



Python+TensorFlow人工智能

该专栏为人工智能入门专栏,采用Python3和TensorFlow实现人工智能相关算法。前期介绍安装流程、基础语法...



¥9.90

去订阅

前面一直强调Python运用到网络爬虫方面非常有效,这篇文章也是结合学习的Python视频知识及我研究生数据挖掘方向的知识,从而简单介绍下Python是如何爬去网络数据的,文章知识非常简单,但是也分享给大家,就当简单入门吧!同时只分享知识,希望大家不要去做破坏网络的知识或侵犯别人的原创文章,主要包括:

- 1.介绍爬取CSDN自己博客文章的简单思想及过程
- 2.实现Python源码爬取新浪韩寒博客的316篇文章

一.爬虫的简单思想

最近看刘兵的《Web数据挖掘》知道,在研究信息抽取问题时主要采用的是三种方法:

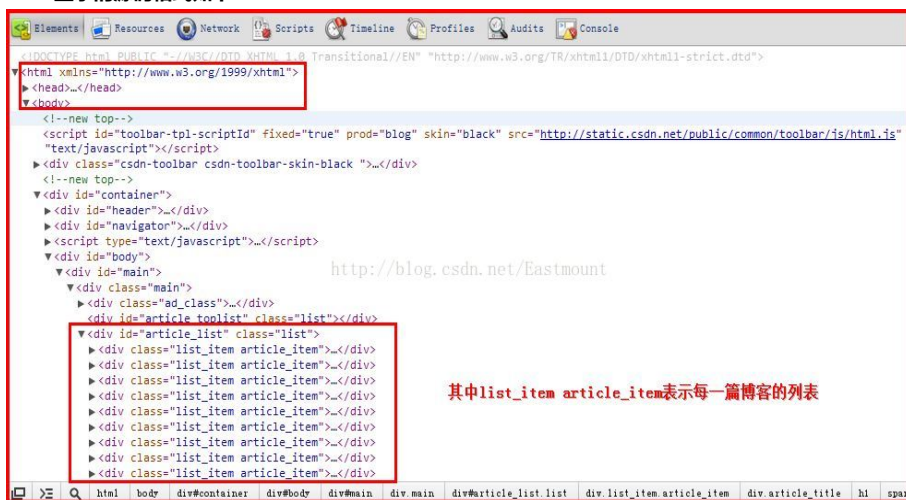
- 1.手工方法: 通过观察网页及源码找出模式,再编写程序抽取目标数据.但该方法无法处理站点数量巨大情形.
- 2.包装器归纳: 它英文名称叫Wrapper Induction,即有监督学习方法,是半自动的.该方法从手工标注的网页或数据记录集中学习一组抽取规则,从而抽取具有类似格式的网页数据.
- 3.自动抽取: 它是无监督方法,给定一张或数张网页,自动从中寻找模式或语法实现数据抽取,由于不需要手工标注,故可以处理大量站点和网页的数据抽取工作.

这里使用的Python网络爬虫就是简单的数据抽取程序,后面我也将陆续研究一些Python+数据挖掘的知识并写这类文章.首先我想获取的是自己的所有CSDN的博客(静态.html文件),具体的思想及实现方式如下:

第一步 分析csdn博客的源码

首先需要实现的是通过分析博客源码获取一篇csdn的文章,在使用IE浏览器按F12或Google Chrome浏览器右键"审查元素"可以分析博客的基本信息.在网页中<http://blog.csdn.net/eastmount>链接了作者所有的博文.

显示的源码格式如下:



其中

原 [Python学习] 专题三.字符串的基础知识

在Python中最重要的数据类型包括字符串、列表、元组和字典等.这篇主要讲述Python的字符串基础知识,包括转义字符串、raw原始字符串、unicode字符串、格式化字符串,及其使用方法和基本操作,基础知识仅分享给大家学习。... 地址: <http://blog.csdn.net/eastmount/article/details/39599061>

2014-09-28 11:10 阅读(144) 评论(0)

第1页 共4页

它的具体html源代码如下:

```
<div class="list_item article_item">
  <div class="article_title">
    <span class="ico ico_type_Original"></span>
    <h1>
      <span class="link_title">
        <a href="/eastmount/article/details/39599061">
          [Python学习] 专题三.字符串的基础知识
        </a>
      </span>
    </h1>
  </div>
  <div class="article_description">
    "
    在Python中最重要的数据类型包括字符串、列表、元组和字典等.该篇主要讲述Python的字符串基础知识.包括转义字符串、raw原始字符串、unicode字符串、格式化字符串,及其使用方法和基本操作,基础知识仅分享与大家学习。...
  </div>
  <div class="article_manage">
    <span class="link_postdate">2014-09-28 11:10</span>
    <span class="link_view" title="阅读次数">...</span>
    <span class="link_comments" title="评论次数">...</span>
  </div>
  <div class="clear"></div>
</div>
<div class="list_item article_item">...</div>
<div class="list_item article_item">...</div>
```

获取一篇博客的链接
包括标题 详情 阅读次数 评论次数

所以我们只需要获取每页中博客<div class="article_title">中的链接,并增加<http://blog.csdn.net>即可.在通过代码:

```
import urllib
content = urllib.urlopen("http://blog.csdn.net/eastmount/article/details/39599061").read()
open('test.html', 'w+').write(content)
```

但是CSDN会禁止这样的行为,服务器禁止爬取站点内容到别人的网上去.我们的博客文章经常被其他网站爬取,但并没有申明原创出处,还请尊重原创.它显示的错误"403 Forbidden".

PS:据说模拟正常上网能实现爬取CSDN内容,读者可以自己去研究,作者此处不介绍.参考(已验证):

<http://www.yihaomen.com/article/python/210.htm>

<http://www.2cto.com/kf/201405/304829.html>

第二步 获取自己所有的文章

这里只讨论思想,假设我们第一篇文章已经获取成功.下面使用Python的find()从上一个获取成功的位置继续查找下一篇文章链接,即可实现获取第一页的所有文章.它一页显示的是20篇文章,最后一页显示剩下的文章.

那么如何获取其他页的文章呢?

原 PHP XAMPP配置PHP环境和Apache80端口被占用解决方案

在使用PHP架构网站时,我们可能会遇到LAMP(Linux+Apache+MySQL+PHP)或WAMP(Windows+Apache+MySQL+PHP)的课程知识,它可以使用XAMPP软件(Apache+MySQL+PHP集成开发包)搭建PHP环境进行网站开发.该文章主要是介绍如何使用XAMPP搭建PHP环境,并解决最常见的一个80端口被系统占用的问题,并实现显示第一个PHP代码程序过程.文章为PHP基础知识,^[1]如果有错误或不足之处见谅!...

2013-09-19 03:06 阅读(1899) 评论(0) 编辑 删除

68条数据 共4页 首页 上一页 1 2 3 4 下一页 尾页

我们可以发现当跳转到不同页时显示的超链接为:

第1页 <http://blog.csdn.net/Eastmount/article/list/1>
第2页 <http://blog.csdn.net/Eastmount/article/list/2>
第3页 <http://blog.csdn.net/Eastmount/article/list/3>
第4页 <http://blog.csdn.net/Eastmount/article/list/4>

这思想就非常简单的,其过程简单如下:

```
for(int i=0;i<4;i++) //获取所有页文章
    for(int j=0;j<20;j++) //获取一页文章 注意最后一篇文章篇数
        GetContent(); //获取一篇文章 主要是获取超链接
```

同时学习通过正则表达式,在获取网页内容图片过程中格外方便.如我前面使用C#和正则表达式获取图片的文章:<http://blog.csdn.net/eastmount/article/details/12235521>

二.爬取新浪博客

上面介绍了爬虫的简单思想,但是由于一些网站服务器禁止获取站点内容,但是新浪一些博客还能实现.这里参照"51CTO学院 智普教育的python视频"获取新浪韩寒的所有博客.

地址为:http://blog.sina.com.cn/s/articlelist_1191258123_0_1.html

采用同上面一样的方式我们可以获取每个<div class="articleCell SG j linedot1">..



此时通过Python获取一篇文章的代码如下:

```
import urllib
content = urllib.urlopen("http://blog.sina.com.cn/s/blog_4701280b0102eo83.html").read()
open('blog.html', 'w+').write(content)
```

可以显示获取的文章,现在需要获取一篇文章的超链接,即:

[《论电影的七个元素》——关于我对电影的一些看法以及《后会无期》的一些消息](http://blog.sina.com.cn/s/blog_4701280b0102eo83.html "《论电影的七个元素》——关于我对电影的一些看法以及《后会无期》的一些消息")

在没有讲述正则表达式之前使用Python人工获取超链接[http,从文章开头查找第一个"<a title",然后接着找到"href="和".html"即可获取](http://blog.sina.com.cn/s/blog_4701280b0102eo83.html)["http://blog.sina.com.cn/s/blog_4701280b0102eo83.html"](http://blog.sina.com.cn/s/blog_4701280b0102eo83.html)。代码如下:

```
#<a title=".." target="_blank" href="http://blog.sina...html">..</a>
#coding:utf-8
con = urllib.urlopen("http://blog.sina.com.cn/s/articlelist_1191258123_0_1.html").read()
title = con.find(r'<a title=')
href = con.find(r'href=',title) #从title位置开始搜索
html = con.find(r'.html',href) #从href位置开始搜索最近html
url = con[href+6:html+5] #href="共6位 .html共5位
print 'url:',url

#输出
url: http://blog.sina.com.cn/s/blog_4701280b0102eohi.html
```

下面按照前面讲述的思想通过两层循环即可实现获取所有文章,具体代码如下:

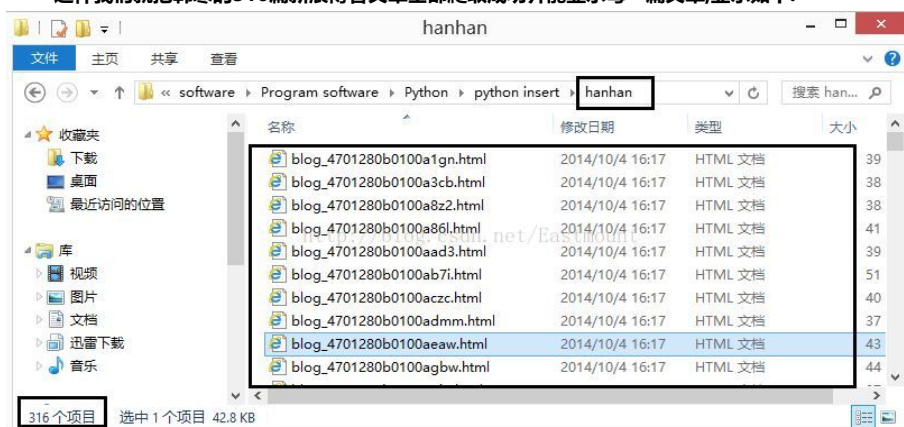
```
#coding:utf-8
import urllib
import time
page=1
while page<=7:
    url=['']*50      #新浪播客每页显示50篇
    temp='http://blog.sina.com.cn/s/articlelist_1191258123_0_'+str(page)+'.html'
    con =urllib.urlopen(temp).read()
    #初始化
    i=0
    title=con.find(r'<a title=')
    href=con.find(r'href='.title)
```

```

html = con.find(r'.html',href)
                                #循环显示文章
while title!=-1 and href!=-1 and html!=-1 and i<50:
    url[i]=con[href+6:html+5]
    print url[i] #显示文章URL
    #下面的从第一篇结束位置开始查找
    title=con.find(r'<a title=',html)
    href=con.find(r'href=',title)
    html = con.find(r'.html',href)
    i=i+1
else:
    print 'end page=',page
#下载获取文章
j=0
while(j<i):
    #前面6页为50篇 最后一页为i篇
    content=urllib.urlopen(url[j]).read()
    open(r'hanhan/'+url[j][-26:], 'w+').write(content) #写方式打开 +表示没有即创建
    j=j+1
    time.sleep(1)
else:
    print 'download'
page=page+1
else:
    print 'all find end'

```

这样我们就把韩寒的316篇新浪博客文章全部爬取成功并能显示每一篇文章,显示如下:



这篇文章主要是简单的介绍了如何使用Python实现爬取网络数据,后面我还将学习一些智能的数据挖掘知识和Python的运用,实现更高效的爬取及获取客户意图和兴趣方面的知识.想实现智能的爬取图片和小说两个软件.

该文章仅提供思想,希望大家尊重别人的原创成果,不要随意爬取别人的文章并没有含原作者信息的转载!最后希望文章对大家有所帮助,初学Python,如果有错误或不足之处,请海涵!

(By:Eastmount 2014-10-4 中午11点 原创CSDN <http://blog.csdn.net/eastmount/>)

参考资料:

1.51CTO学院 智普教育的python视频http://edu.51cto.com/course/course_id-581.html

2.《Web数据挖掘》刘兵著

👍 点赞 6 ☆ 收藏 ➦ 分享



Eastmount 博客专家

发布了444 篇原创文章 · 获赞 5918 · 访问量 484万+