链滴

python 爬虫（SGMLParser）

作者：cactus0509

原文链接：https://ld246.com/article/1470105137836

来源网站：链滴

`# -- coding: utf-8 --
import requests
import urllib2
import sys
from sgmllib import SGMLParser
from cgitb import text
reload(sys)
sys.setdefaultencoding('utf8')
class CLAS_EXPERT_LIST(SGMLParser):
def init(self):
reload(sys)
SGMLParser.init(self)
self.is_a = ""
self.name = []
self.urls = []
<pre><code class="highlight-chroma"><span class="highlight
cl">def start_a(self, attrs):
    for k, v in attrs :
        if k=='href'
nd v.count('detail') > 0 :
            self.is_a =

            self.urls.ap
end(v)
def end_a(self):
    self.is_a = 0

def handle_data(s
lf, text):
    if self.is_a == 1:
        self.name.ap
end(text)
</code></pre>
class EXPERT(SGMLParser):
<pre><code class="highlight-chroma"><span class="highlight
cl">def __init__(self):
    SGMLParser.__in
t__(self)
    self.is_div = 0
    self.is_h3 = 0
    self.is_div_p = 0

    self.p_cnt = 0
    self.image = {}

    self.is_div_exper
 = 0
    self.is_div_exper
_p = 0
    self.is_div_exper
_p_cnt = 0

def start_div(self,

```
ttrs):
    for k, v in attrs :
        if k=='class' and v.count('name') &gt; 0  :
            self.is_div = 1
        if k=='class' and v.count('expert_content') &gt; 0  :
            self.is_div_expert = 1

def end_div(self):
    if self.is_div == :
        self.is_div = 0
    elif self.is_div_expert == 1 :
        self.is_div_expert = 0

def start_h3(self, attrs):
    if self.is_div :
        self.is_h3 = 1

def end_h3(self):
    self.is_h3 = 0

def start_p(self, atts):
    if self.is_div == :
        self.is_div_p = 1
    elif self.is_div_expert == 1 :
        self.is_div_expert_p = 1
        self.is_div_expert_p_cnt = self.is_div_expert_p_cnt + 1

def end_p(self):
    if self.is_div :
        self.is_div_p = 0
        self.p_cnt = self.p_cnt + 1
    elif self.is_div_expert == 1 :
        self.is_div_expert_p = 0
```

```
</span></span><span class="highlight-line"><span class="highlight-cl">def handle_data(self, text):
</span></span><span class="highlight-line"><span class="highlight-cl">    try:
</span></span><span class="highlight-line"><span class="highlight-cl">        if self.is_div = 1:
</span></span><span class="highlight-line"><span class="highlight-cl">            if self.is_h3 == 1 :
</span></span><span class="highlight-line"><span class="highlight-cl">                self.image["name"] = text
</span></span><span class="highlight-line"><span class="highlight-cl">            if self.is_di_p == 1 :
</span></span><span class="highlight-line"><span class="highlight-cl">                if self.p_nt == 0 :
</span></span><span class="highlight-line"><span class="highlight-cl">                    self.image["job"] = text
</span></span><span class="highlight-line"><span class="highlight-cl">                else:
</span></span><span class="highlight-line"><span class="highlight-cl">                    self.image["title"] = text
</span></span><span class="highlight-line"><span class="highlight-cl">        if self.is_div_epert  == 1:
</span></span><span class="highlight-line"><span class="highlight-cl">            if self.is_di_expert_p == 1 :
</span></span><span class="highlight-line"><span class="highlight-cl">                #print self.is_div_expert_p_cnt,text
</span></span><span class="highlight-line"><span class="highlight-cl">                if self.is_div_expert_p_cnt == 2:
</span></span><span class="highlight-line"><span class="highlight-cl">                    self.image["employer"] = text
</span></span><span class="highlight-line"><span class="highlight-cl">                elif self.i_div_expert_p_cnt == 6:
</span></span><span class="highlight-line"><span class="highlight-cl">                    self.image["filed"] = text
</span></span><span class="highlight-line"><span class="highlight-cl">                elif self.i_div_expert_p_cnt == 16:
</span></span><span class="highlight-line"><span class="highlight-cl">                    self.image["conn_info"] = text
</span></span><span class="highlight-line"><span class="highlight-cl">
</span></span><span class="highlight-line"><span class="highlight-cl">            #print text
</span></span><span class="highlight-line"><span class="highlight-cl">            #print self.mage["name"] ,  self.image["title"] , self.image["job"] , self.image["employer"] , self.image["file"] , self.image["conn_info"]
</span></span><span class="highlight-line"><span class="highlight-cl">
</span></span><span class="highlight-line"><span class="highlight-cl">
</span></span><span class="highlight-line"><span class="highlight-cl">    except Exceptio,e:
</span></span><span class="highlight-line"><span class="highlight-cl">        print e
</span></span></code></pre>
<p>def list_expert():<br>
headers = {<br>
"Connection": "keep-alive",<br>
"Cookie":"Ecp_IpLoginFail=160726111.205.187.18; kc_cnki_net_uid=ff38e944-e46c-2d76-349c24a97e03ded8; ASP.NET_SessionId=ysbae4exnu0vkugigsdnknps; AutoIpLogin=; LID=; SID=12103; CNZZDATA4922505=cnzz_eid%3D1343153553-1469773415-%26ntime%3D1469782211
 FileNameM=cnki%3A; c_m_LinID=LinID=WEEvREcwSlJHSldTTGJhYlRtMVNwOTZ6Q1UzaHd
```

OFN2RzR2MEEyUkJPWmE=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4IQMovwHtwkF4VYPoH
KxJw!!&ot=07/29/2016 18:16:04",
"Host":"elib.cnki.net",
"Accept":"text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8",
"Accept-Encoding":"gzip, deflate",
"Accept-Language":"zh-CN,zh;q=0.8,en-US;q=0.5,en;q=0.3",
"Referer":"<a href="https://ld246.com/forward?goto=http%3A%2F%2Fwww.example.com%2
%2522" target="_blank" rel="nofollow ugc">http://www.example.com/",
"User-Agent":" Mozilla/5.0 (Macintosh; Intel Mac OS X 10.11; rv:47.0) Gecko/20100101 Firefox
47.0"
}
r = requests.get('<a href="https://ld246.com/forward?goto=http%3A%2F%2Fwww.chinathink
anks.org.cn%2Fcontent%2Fexpert" target="_blank" rel="nofollow ugc">http://www.chinathin
tanks.org.cn/content/expert')
content=r.text
listname = CLAS_EXPERT_LIST()
listname.feed(content)
rn = zip(listname.urls,listname.name)
return rn
def get_expert(url,name):
headers = {
"Connection": "keep-alive",
"Cookie":"Ecp_IpLoginFail=160726111.205.187.18; kc_cnki_net_uid=ff38e944-e46c-2d76-349c
24a97e03ded8; ASP.NET_SessionId=ysbae4exnu0vkugigsdnknps; AutoIpLogin=; LID=; SID=1
2103; CNZZDATA4922505=cnzz_eid%3D1343153553-1469773415-%26ntime%3D1469782211
FileNameM=cnki%3A; c_m_LinID=LinID=WEEvREcwSlJHSldTTGJhYlRtMVNwOTZ6Q1UzaHd
OFN2RzR2MEEyUkJPWmE=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4IQMovwHtwkF4VYPoH
KxJw!!&ot=07/29/2016 18:16:04",
"Host":"elib.cnki.net",
"Accept":"text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8",
"Accept-Encoding":"gzip, deflate",
"Accept-Language":"utf-8,zh;q=0.8,en-US;q=0.5,en;q=0.3",
"Referer":"<a href="https://ld246.com/forward?goto=http%3A%2F%2Fwww.example.com%2
%2522" target="_blank" rel="nofollow ugc">http://www.example.com/",
"User-Agent":" Mozilla/5.0 (Macintosh; Intel Mac OS X 10.11; rv:47.0) Gecko/20100101 Firefox
47.0"
}
r = requests.get(url)
content=r.text
expert = EXPERT()
expert.feed(content)
return expert.image
if name == "main":
try:
i = 0
import chardet
with open("/tmp/expert.txt","w") as f:
exports = list_expert()
for k,v in exports:
v = get_expert(k,v)
name,job,title,employer,filed,conn_info=None,None,None,None,None,None
if "name" in v:
name= v["name"]
if "job" in v:

job= v["job"]
if "title" in v:
title= v["title"]
if "employer" in v:
employer= v["employer"]
if "filed" in v:
filed= v["filed"]
if "conn_info" in v:
conn_info= v["conn_info"]</p>
<pre><code class="highlight-chroma"><span class="highlight
cl">   print "{0}#{1}#{2}#{3}#{4}#{5}".format(name,job,title,employer,filed,conn_info)
            f.write("{0}
{1}#{2}#{3}#{4}#{5}\n".format(name,job,title,employer,filed,conn_info))
            f.flush()
            i = i + 1
            left = div
od(i,50)
            if left[1]
= 0 :
                print i

except Exception,e
:
    print e
`
</code></pre>