



Python与金融数据挖掘(7)

文欣秀

wenxinxiu@ecust.edu.cn

各个网站的Robots协议


网站名称	Robtos协议
百度	https://www.baidu.com/robots.txt
京东	https://www.jd.com/robots.txt
QQ	https://www.qq.com/robots.txt
人民网	http://www.people.com.cn/robots.txt

案例分析


人民网_网上的人民日报 people.com.cn

《求是》杂志发表习近平总书记重要文章《加强文化遗产保护传承 弘扬中华优秀传统文化》

中华文明源远流长，从未中断，塑造了我们伟大的民族。要让文物说话，让历史说话，让文化说话。世界由丰富多彩的文明构成，中国是有着世界上最古老历史和文



中邮理财
PSBC WEALTH MANAGEMENT



超1300万户
累计服务客户

超1000亿元
累计创造收益

构建多元理财产品体系，为您的财富保驾护航
中邮理财：为您的财富保驾护航

数据来源：中邮理财。数据截止日期：2024年2月29日。

时习之|习近平要求抓好这件“头
为强国建设、民族复兴提供坚强安
以金融高质量发展助力强国建设、
我国连续7年保持货物贸易第一大
政策问答|如何提高产业和就业帮
我国创新发展农村客货邮融合：乡
任平：以改革创新为根本动力——从新能
国家安全机关公布多起危害我国国
现代职业教育体系加快构建 各地
汇聚合力务实举措 为人工智能发展
上海连续三年推出的“数字化转型
第39届“上海之春”落幕，让年轻

简单爬虫程序基本步骤

- ◆ 获取网页源码
- ◆ 根据源码中所在链接特点写出正则表达式
- ◆ 用正则对象匹配获取目标链接
- ◆ 用循环结构遍历目标链接并自动下载信息

Requests对象的属性

属性	说明
r. text	网页响应内容的 字符串 形式
r. encoding	猜测 网页响应内容编码方式
r. apparent_encoding	从网页内容 分析 出编码方式
r. content	网页响应内容的 二进制 形式
r.status_code	200 表示连接 成功 ，404表示失败

爬取单张图片

```
import requests
url="http://www.people.com.cn/NMediaFile/2023/0329/MAIN202303
291337317271860541036.jpg"
r=requests.get(url)
data=r.content
fobj=open("result.jpg","wb")
fobj.write(data)
fobj.close()
```

爬取网页

```
import requests
url="http://www.people.com.cn"
r=requests.get(url)
print(r.status_code)
r.encoding=r.apparent_encoding
print(r.encoding)
data=r.text
with open("paper.txt","w",encoding="GB2312") as fobj:
    fobj.write(data)
```

案例分析

“C:\素材”文件夹中h.txt为已爬取的某新闻网站的静态html文本文件（编码格式为UTF-8），其中新闻链接和标题的呈现特点是“标题”，请利用正则方法，筛选其中新闻链接和标题，保存在C:\KS\news.csv（编码格式为GBK，结果示例如图），程序保存在C:\KS目录下，名为 4_5.py。

1	http://www.ce.cn/xwzx/gnsz/gdxw/202007/24/t20200724_35386699.shtml	中国经济增长重开马力		
2	http://www.ce.cn/xwzx/gnsz/gdxw/202007/24/t20200724_35386703.shtml	外贸有所回稳		
3	http://news.cctv.com/2020/07/23/ARTIOgdNR4557PJua229m827200723.shtml	2季度货运量正增长		
4	http://news.cctv.com/2020/07/23/ARTluXNPwTxxLI1IRPlz8sjp200723.shtml	工业生产稳步回升		
5	https://wap.peopleapp.com/article/5760724/5682238	王小良返乡脱贫记		
6	https://3w.huanqiu.com/a/de583b/3zAz3PASXzE?agt=8	"天问一号"中国首次火星探测这四大看点请收好		
7	https://wap.gmdaily.cn/article/pb517fcb032b24c93ae0cc43f3d24f44f	遵义:脱贫到致富		
8	https://xhpfmapi.zhongguowangshi.com/vh512/share/9272522	乌蒙"同心"战贫困		
9	https://wap.gmdaily.cn/article/pfe06986dc832498ba3e1dc7c3c7ae3c4	古胜村逆袭记		
10	https://xhpfmapi.zhongguowangshi.com/vh512/share/9272405	甩穷帽子吃"生态饭"		
11	http://news.cctv.com/2020/07/23/ARTIp20aylqkJecYvJYprLgz200723.shtml	眉山果园村"幸福密码"		
12	http://www.ce.cn/xwzx/gnsz/gdxw/202007/23/t20200723_35381001.shtml	"码上黔行"		
13	https://mp.weixin.qq.com/s/rei6k_eVgZO7yRujrsVxA	关于有不法分子冒用中央网信办举报中心名义开展		
14	https://3w.huanqiu.com/a/438198/3zBTsZ7tF7X?agt=8	外交部:美国驻成都总领事馆人员从事与身份不符		
15	https://baijiahao.baidu.com/s?id=1673081065892684209&wfr=content	央视实拍美驻成都总领馆:门前市民熙攘 院内情		

re库的内置函数

- ◆ `match()`: 用于从起始位置匹配
- ◆ `search()`: 搜索整个字符串, 返回第一次出现位置
- ◆ `findall()`: 以列表形式返回全部能匹配的字符串
- ◆ `compile()`: 创建一个正则表达式对象

内置函数示例

```
>>> import re
```

```
>>> string="I love it. It loves me."
```

```
>>> result=re.findall("it", string)
```

```
>>> print(result)
```

问题： 如何实现不区分大小写？

正则表达式修饰符含义

修饰符	描述
re.I	使匹配对大小写不敏感
re.L	做本地化识别 (locale-aware) 匹配
re.M	多行匹配, 影响 ^ 和 \$
re.S	使 . 匹配包括换行在内的所有字符
re.U	根据Unicode字符集解析字符。这个标志影响 \w, \W, \b, \B.
re.X	该标志通过给予你更灵活的格式以便你将正则表达式写得更易于理解。

内置函数示例

```
>>> import re
>>> string="I love it. It loves me."
>>> test=re.compile('it', re.I)
>>> result=re.findall(test, string) #result=test.findall(string)
>>> print(result)
```

问题： 如何实现查找所有以e结尾的单词？

字符串匹配表

模式	描述
\d	匹配一个数字字符
\w	匹配一个字母、数字及下划线字符
\s	匹配一个空白字符
.	匹配一个任意字符，换行符除外
\n	匹配一个换行符
\t	匹配一个制表符

字符串匹配表

模式	描述
*	匹配前面的字符0次或n次
+	匹配前面的字符1次或n次
?	匹配前面的字符0次或1次
()	匹配括号内表达式，也表示一个组
{m, n}	匹配m-n个字符
[]	表示字符范围，方括号中只能取一个

内置函数示例

```
import re  
  
string="I love it. It loves me."  
  
result=re.findall("\w{1,4}e", string)  
  
print(result)
```

`['love', 'love', 'me']`

正则表达式示例

```
import re
s="<img src=\"C:\\XH.jpg\" width=\"300\"/>
  <img src=\"C:\\FX.jpg\" width=\"300\"/>"
result=re.findall('<img src=\"(.*)\"', s)
print(result)
```

['C:\\XH.jpg', 'C:\\FX.jpg']

正则表达式示例

```
import re
s="<img src=\"C:\\XH.jpg\" width=\"300\"/>
  <img src=\"C:\\FX.jpg\" width=\"300\"/>"
result=re.findall('<img src=\"(.*)\" width=\"(.*)\"', s)
print(result)
```

[('C:\\XH.jpg', '300'), ('C:\\FX.jpg', '300')]

豌豆荚正则示例

```
page=u"  
<li class="parent-cate">  
<a class="cate-link" c>旅游出行</a><ul>  
<li class="child-cate"><a href="http://www.wandoujia.com/category/596"  
title="综合旅游服务">综合旅游服务</a></li>  
<li class="child-cate"><a href="http://www.wandoujia.com/category/598"  
title="攻略">攻略</a></li>  
<li class="child-cate"><a href="http://www.wandoujia.com/category/600"  
title="酒店·住宿">酒店·住宿</a></li>  
</ul></li>  
"
```

豌豆荚正则示例

```
#coding=utf-8
import re
page=u"<li class='parent-cate'>
<a class='cate-link' c>旅游出行</a>
<ul><li class='child-cate'><a href='http://www.wandoujia.com/category/596' title='综合旅游服务'>
综合旅游服务</a></li>
<li class='child-cate'><a href='http://www.wandoujia.com/category/598' title='攻略'>攻略</a></li>
<li class='child-cate'><a href='http://www.wandoujia.com/category/600' title='酒店·住宿'>酒店·住
宿</a></li>
</ul></li>
'"

data=re.findall('href="(*?)"' title="(*?)"', page)
print(data)
```

豌豆荚正则存入文件

```
#coding=utf-8
import re
page=u" <li class="parent-cate">
<a class="cate-link" c>旅游出行</a>
<ul><li class="child-cate"><a href="http://www.wandoujia.com/category/596" title="综合旅游服务">综合旅游服务</a></li>
<li class="child-cate"><a href="http://www.wandoujia.com/category/598" title="攻略">攻略</a></li>
<li class="child-cate"><a href="http://www.wandoujia.com/category/600" title="酒店·住宿">酒店·住宿</a></li>
</ul></li>
""

data=re.findall('href="(.*?)" title="(.*?)"', page)
fobj=open("test.txt", "w")
for line in data:
    fobj. write(line[0]+" "+line[1]+"\\n")
fobj. close()
```

案例分析

“C:\素材”文件夹中h.txt为已爬取的某新闻网站的静态html文本文件（编码格式为UTF-8），其中新闻链接和标题的呈现特点是“标题”，请利用正则方法，筛选其中新闻链接和标题，保存在C:\KS\news.csv（编码格式为GBK，结果示例如图），程序保存在C:\KS目录下，名为 4_5.py。

1	http://www.ce.cn/xwzx/gnsz/gdxw/202007/24/t20200724_35386699.shtml	中国经济增长重开马力		
2	http://www.ce.cn/xwzx/gnsz/gdxw/202007/24/t20200724_35386703.shtml	外贸有所回稳		
3	http://news.cctv.com/2020/07/23/ARTIOgdNR4557PJua229m827200723.shtml	2季度货运量正增长		
4	http://news.cctv.com/2020/07/23/ARTIuXNPwTxzLI1IRPlz8sjp200723.shtml	工业生产稳步回升		
5	https://wap.peopleapp.com/article/5760724/5682238	王小良返乡脱贫记		
6	https://3w.huanqiu.com/a/de583b/3zAz3PASXzE?agt=8	"天问一号"中国首次火星探测这四大看点请收好		
7	https://wap.gmdaily.cn/article/pb517fcb032b24c93ae0cc43f3d24f44f	遵义:脱贫到致富		
8	https://xhpfmapi.zhongguowangshi.com/vh512/share/9272522	乌蒙"同心"战贫困		
9	https://wap.gmdaily.cn/article/pfe06986dc832498ba3e1dc7c3c7ae3c4	古胜村逆袭记		
10	https://xhpfmapi.zhongguowangshi.com/vh512/share/9272405	甩穷帽子吃"生态饭"		
11	http://news.cctv.com/2020/07/23/ARTIp20aylqkJecYvJYprLgz200723.shtml	眉山果园村"幸福密码"		
12	http://www.ce.cn/xwzx/gnsz/gdxw/202007/23/t20200723_35381001.shtml	"码上黔行"		
13	https://mp.weixin.qq.com/s/rei6k_eVgZO7yRujrsVxA	关于有不法分子冒用中央网信办举报中心名义开展		
14	https://3w.huanqiu.com/a/438198/3zBTsZ7tF7X?agt=8	外交部:美国驻成都总领事馆人员从事与身份不符		
15	https://baijiahao.baidu.com/s?id=1673081065892684209&wfr=content	央视实拍美驻成都总领馆: 门前市民熙攘 院内情		

(.*?)

案例分析

```
import re
fobj=open("C:\\素材\\h.txt","r")
paper=fobj.read()

reg='<a href="(http://.*?)" mon="ct=1&a=2&c=top&pn=\d{1,2}"
target="_blank">(.*?)</a>'
result=re.findall(reg,paper)

fobj=open("C:\\KS\\news.csv", "w")
for line in result:
    fobj. write(line[0]+"," +line[1]+"\n")
fobj. close()
```

爬取人民网链接和标题

```
import requests
import re
url="http://www.people.com.cn"
html=requests.get(url)
html.encoding=html.apparent_encoding
data=html.text
#print(data)
```

爬取人民网链接和标题

```
reg='<li><a href="(.*?)" target="_blank">(.*?)</a></li>'  
urls=re.findall(reg, data)  
print(urls)  
fobj=open("result.csv",'w', encoding="gb2312")  
for titu in urls:  
    fobj.write(titu[0]+", "+titu[1]+'\\n')  
fobj.close()
```


爬取新闻图片素材

```
import requests
import re
url='http://www.people.com.cn'
r=requests.get(url)
r.encoding=r.apparent_encoding
paper=r.text
cond='>> res=' 华能信托本年实现利润32.05亿元 '
```

```
>>> res=res.strip()
```

```
>>> res
```

```
'华能信托本年实现利润32.05亿元'
```

数据清洗常见方法

◆ 用split()函数截取需要的内容

```
>>> date='2019-01-20 10:10:10'
```

```
>>> date=date.split(' ')[0]
```

```
>>> date    '2019-01-20'
```

数据清洗常见方法

◆ 用sub()函数进行内容替换

短语标签, 用来呈现为被强调的文本

```
>>> import re
```

```
>>> title='阿里<em>巴巴</em>人工智能再发力'
```

```
>>> title=re.sub('<.*?>', '', title)
```

```
>>> title    '阿里巴巴人工智能再发力'
```

思考

```
import re
res = '<h3 class="c-title"><a href="网址" data-click="{一堆英文}"><em>阿里巴巴</em>代码竞赛现全球首位AI评委 能为代码质量打分</a>'
p_title = '<h3 class="c-title">.*?>(.*?)</a>'
title = re.findall(p_title, res)
print(title)
```

['阿里巴巴代码竞赛现全球首位AI评委 能为代码质量打分']

解决方案

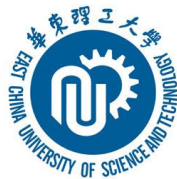
```
import re

title = ['<em>阿里巴巴</em>代码竞赛现全球首位AI评委
能为代码质量打分']

result = re.sub('<.*?>', "", title[0])

print(result)
```

阿里巴巴代码竞赛现全球首位AI评委 能为代码质量打分



谢 谢