[Pandas 读写html](https://geek-docs.com/pandas/pandas-read-write/pandas-speaking-reading-and-writing-html.html)

**Pandas 读写html**，pandas提供read\_html(),to\_html()两个函数用于读写html格式的文件。这两个函数非常有用，把DataFrame等复杂的数据结构转换成HTML表格很简单，无需编写一长串HTML代码就能实现。pandas这方面的能力很强大，如果你从事web开发，这个功能将给你带来很多便捷。

读取网页数据这种操作被称为**网页抓取**，应用比较广泛，它逐渐演变成数据分析过程中的一项基础操作，被整合到了数据分析的第一步-数据挖掘和数据准备。

**写入数据到HTML文件**

现在我们来学习把DataFrame转换成HTML表格的方法。DataFrame的内部结构被自动转换为嵌入在表格中的<TH>,<TR>,<TD>标签，保留所有内部层级结构。使用该函数，无需了解HTML知识。因为有时候DataFrame等数据结构太复制，规模很大，所以对需要开发网页的人来说，往HTML文件中写入数据的函数作用很大。

to\_html()函数可以直接把DataFrame转换成HTML表格，该函数在Pandas数据结构内部定义，因为可以直接在DataFrame对象上调用to\_html()函数，如下例所示：

import pandas as pd

import numpy as np

frame = pd.DataFrame(pd.DataFrame(np.arange(4).reshape(2,2)))

print(frame)

print(frame.to\_html())

Python

Copy

输出结果如下所示:

0 1

0 0 1

1 2 3

<table border="1" class="dataframe">

<thead>

<tr style="text-align: right;">

<th></th>

<th>0</th>

<th>1</th>

</tr>

</thead>

<tbody>

<tr>

<th>0</th>

<td>0</td>

<td>1</td>

</tr>

<tr>

<th>1</th>

<td>2</td>

<td>3</td>

</tr>

</tbody>

</table>

Shell

Copy

如上所示，该函数按照DataFrame的内部结构，正确生成了创建HTML表格所需的HTML标签。

下面的例子演示如何在HTML文件中自动生成表格，我们创建一个比上面更加复杂、具有索引和列名称的DataFrame对象。如下所示:

import pandas as pd

import numpy as np

frame = pd.DataFrame(pd.DataFrame(np.random.random((4,4)),

index=['white', 'black', 'red', 'blue'],

columns=['up', 'down', 'right', 'left']))

print(frame)

Python

Copy

输出结果如下:

up down right left

white 0.085296 0.447792 0.366266 0.534075

black 0.408325 0.461076 0.713888 0.108599

red 0.851112 0.664453 0.596685 0.704459

blue 0.365703 0.917268 0.595423 0.590646

Shell

Copy

现在，请把注意力放在如何生成一个字符串并把它写入到HTML页面上，这个例子虽然短小，但是可以帮助快速理解和测试pandas的功能。下面定义HTML的页面内容：

s = ['<HTML>']

s.append('<HEAD><TITLE>极客教程</TITLE></HEAD>')

s.append('<BODY>')

s.append(frame.to\_html())

s.append('</BODY></HTML>')

html = ''.join(s)

Python

Copy

将HTML页面内容html写入到文件myframe.html中：

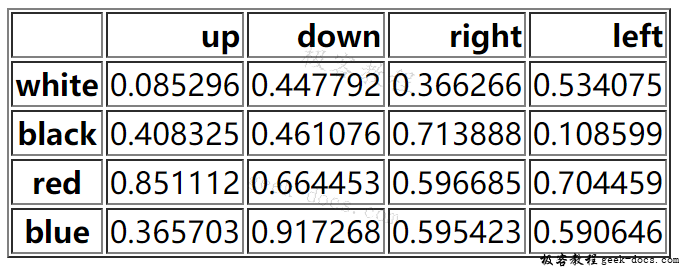
html\_file = open('myframe.html', 'w')

html\_file.write(html)

html\_file.close()

Python

Copy

运行上面代码后，工作目录中多了myframe.html文件，使用web浏览器打开它，显示内容如下:  


**从HTML文件读取数据**

如上所示，Pandas可以直接用DataFrame生成HTML表格，同样可以读取HTML文件。read\_html()函数解析HTML页面，寻找HTML表格。如果找到，就将其转换为可以直接用于数据分析的DataFrame对象。

即使只有一个表格，read\_html()函数也会返回一个DataFrame列表。至于要解析的数据源，可是支持多种数据类型。下面解析上例中创建的HTML文件：

import pandas as pd

web\_frames = pd.read\_html("myframe.html")

print(web\_frames[0])

Python

Copy

输出结果如下:

Unnamed: 0 up down right left

0 white 0.085296 0.447792 0.366266 0.534075

1 black 0.408325 0.461076 0.713888 0.108599

2 red 0.851112 0.664453 0.596685 0.704459

3 blue 0.365703 0.917268 0.595423 0.590646

Shell

Copy

如上所示，所有与HTML表格无关的标签都没有考虑在内。

read\_html()函数最常用的模式是以网址为参数，直接解析并抽取网页中的表格。http://www.meccanismocomplesso.org/meccanismo-complesso-sito-2/classifica-punteggio/网站存在表格如图所示：  
  
解析并抽取网页的表格，如下所示:

import pandas as pd

web\_frames = pd.read\_html("http://www.meccanismocomplesso.org/meccanismo-complesso-sito-2/classifica-punteggio/")

print(web\_frames[0])

输出结果如下所示:

# Nome Exp Livelli

0 1 Fabio Nelli 17850 NaN

1 2 admin 9029 NaN

2 3 BrunoOrsini 2168 NaN

3 4 Berserker 700 NaN

.. ... ... ... ...

246 247 mario.testini@alice.it 4 NaN

247 248 Armote 4 NaN

248 249 Fabri 4 NaN

249 250 pierr86 4 NaN