

Pandas爬虫

讲师：萨缪尔 Samuel



萨缪尔老师

网易数据分析教研负责人;

前盛大游戏战略规划总监、前腾讯游戏商业智能中心Leader

上海交通大学高金硕士、《哈佛管理导师》外部导师

- 知乎大V：「萨缪尔」主要聚集与商业分析、行业与战略分析、Python数据分析等
- BAT互联网巨头商业洞察分析 & 咨询公司战略咨询背景
- 擅长行业趋势研究和战略管理咨询工具，为20多家上市公司提供战略发展决策建议

课程

亮点

1

爬虫的步骤

2

Pandas爬虫说明

3

案例分析

爬虫的步骤



1

爬虫获取数据的四个步骤

Step 1

分析网页URL

打开网站

静态网页

动态网页

Step 2

请求网页数据

请求网页数据

发送get请求

定制请求头

Step 3

解析网页数据

Beautiful Soup

XPath

re正则表达式

Step 4

存储网页数据

结构化文本

txt、csv、json或excel等

数据库储存

如MySQL、MongoDB或SQLite

Pandas爬虫说明



pandas.read_html(url)

```
pandas.read_html( io, match='.+', flavor=None, header=None, index_col=None,
skiprows=None, attrs=None, parse_dates=False, thousands=', ', encoding=None,
decimal='.', converters=None, na_values=None, keep_default_na=True,
displayed_only=True)
```

pandas.read_html(url)

io: str, path object 或 file-like object URL, file-like对象或包含HTML的原始字符串。请注意, lxml仅接受http, ftp和文件url协议。如果网址以'https'您可以尝试删除's'。

match: str 或 compiled regular expression, 可选参数将返回包含与该正则表达式或字符串匹配的文本的表集。除非HTML非常简单, 否则您可能需要在此处传递非空字符串。默认为 "" (匹配任何非空字符串)。默认值将返回页面上包含的所有表。此值转换为正则表达式, 以便Beautiful Soup和lxml之间具有一致的行为。

header: int 或 list-like 或 None, 可选参数该行(或MultiIndex)用于创建列标题。

index_col: int 或 list-like 或 None, 可选参数用于创建索引的列(或列列表)。

skiprows: int 或 list-like 或 slice 或 None, 可选参数解析列整数后要跳过的行数。从0开始。如果给出整数序列或切片, 将跳过该序列索引的行。请注意, 单个元素序列的意思是“跳过第n行”, 而整数的意思是“跳过n行”。

attrs: dict 或 None, 可选参数这是属性的词典, 您可以传递该属性以用于标识HTML中的表。在传递给lxml或Beautiful Soup之前, 不会检查它们的有效性。但是, 这些属性必须是有效的HTML表属性才能正常工作。例如, attrs = {'id': 'table'} 是有效的属性字典, 因为 'id' HTML标记属性是任何HTML标记的有效HTML属性, 这个文件。attrs = {'asdf': 'table'} 不是有效的属性字典, 因为 'asdf' 即使是有效的XML属性, 也不是有效的HTML属性。可以找到有效的HTML 4.01表属性这里。可以找到HTML 5规范的工作草案这里。它包含有关现代Web表属性的最新信息。

案例分析



操作案例1：爬取单网页GDP的数据

In [85]: #单个网页实现，爬取GDP数据

```
import pandas as pd

df=[]
url='https://www.kylc.com/stats/global/yearly_overview/g_gdp.html'
data=pd.read_html(url)[0]
data=pd.DataFrame(data)

data
```

In [85]:

	排名	国家/地区	所在洲	年份	GDP(美元)	占世界比重
0	NaN	全世界	NaN	2020	84.71万亿 (84,705,425,882,119)	NaN
1	1	美国	美洲	2020	20.94万亿 (20,936,600,000,000)	24.7170%
2	NaN	欧盟地区	NaN	2020	15.19万亿 (15,192,652,399,779)	17.9359%
3	2	中国	亚洲	2020	14.72万亿 (14,722,730,697,890)	17.3811%
4	3	日本	亚洲	2019	5.06万亿 (5,064,872,875,604)	NaN
...
204	200	帕劳	大洋洲	2019	2.68亿 (268,354,900)	NaN
205	201	马绍尔群岛	大洋洲	2019	2.39亿 (239,462,200)	NaN
206	202	基里巴斯	大洋洲	2020	2.0亿 (199,573,325)	0.0002%
207	203	瑙鲁	大洋洲	2019	1.18亿 (118,223,430)	NaN
208	204	图瓦卢	大洋洲	2020	4886万 (48,855,550)	0.0001%

209 rows x 6 columns

操作案例2：连续多年爬虫+数据清洗+文件保存

- 步骤1：pandas通过for循环爬取1960-2020年的数据
- 步骤2：数据清洗（删除重复行、格式转换等）
- 步骤3：把数据从文本清洗为数据
- 步骤4：格式调整并保存到CSV文件里
- 步骤5：数据分析（本节课暂时不讲）

课程总结

- 1 爬虫的4个基本步骤：分析URL、请求数据、解析数据、存储数据
- 2 Pandas爬虫的函数说明： `Pandas.read_html(url)`
- 3 通过案例深入了解pandas爬虫

谢谢观看

