

Pandas文本处理

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

文本处理的主要方式

2

案例讲解

文本处理的主要方式



基本的文本操作

功能	函数
拆分或替换字符串	split、replace
拼接	cat
提取子字符串	extract

案例



2

替换函数str.replace 及案例

```
import sys

▼ for i in range(0,10079):
▼     if str(dts.iloc[i,7]).find('(')>=0:
        dts.iloc[i,7]=str(re.findall('\((.*)\)',str(dts.iloc[i,7])))
▼     else:
        dts.iloc[i,7]=str(dts.iloc[i,7].replace(',',''))

#替换掉其中的符号
▼ for i in range(0,10079):
    dts.iloc[i,7]=dts.iloc[i,7].replace(',','').replace('\','').replace('[','').replace(']',')')

#转化为int数据类型
dts['GDP(十亿美元)']=dts['GDP(十亿美元)'].astype(int)
dts['GDP(十亿美元)']=round(dts['GDP(十亿美元)']/1000000000,1)
dts
```

Unnamed: 0	排名	国家/地区	所在洲	GDP(美元)	占世界%	Year	GDP(十亿美元)	序列	
0	0	NaN	全世界	NaN	1.38万亿 (1,384,628,173,213)	NaN	1960	1384.6	1
1	1	1.0	美国	美洲	5433.0亿 (543,300,000,000)	39.2380%	1960	543.3	2
2	2	NaN	欧盟地区	NaN	3644.48亿 (364,448,431,697)	26.3210%	1960	364.4	3
3	3	2.0	英国	欧洲	732.34亿 (73,233,967,692)	5.2891%	1960	73.2	4

课程总结



1

文本处理的几种方式：拆分、替换、提取、合并

2

Replace案例讲解

谢谢观看

