

pandas

October 11, 2023

Author: lcdse7en

Email: 2353442022@qq.com

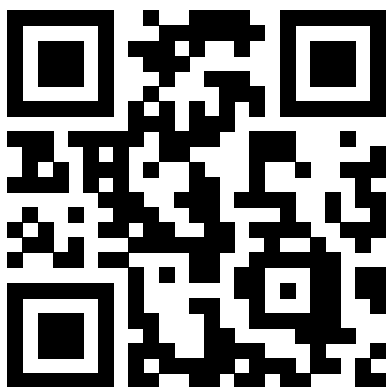
摘要

Pandas是Python的核心数据分析支持库，提供了快速、灵活、明确的数据结构，旨在简单、直观地处理关系型、标记型数据。Pandas的目标是成为 Python 数据分析实践与实战的必备高级工具，其长远目标是成为最强大、最灵活、可以支持任何语言的开源数据分析工具。经过多年不懈的努力，Pandas离这个目标已经越来越近了。

Pandas的主要数据结构是Series（一维数据）与DataFrame（二维数据），这两种数据结构足以处理金融、统计、社会科学、工程等领域里的大多数典型用例。对于R用户，DataFrame提供了比R语言data.frame更丰富的功能。Pandas基于NumPy开发，可以与其它第三方科学计算支持库完美集成。

Pandas 就像一把万能瑞士军刀，下面仅列出了它的部分优势：

- 处理浮点与非浮点数据里的缺失数据，表示为 NaN；
- 大小可变：插入或删除 DataFrame 等多维对象的列；
- 自动、显式数据对齐：显式地将对象与一组标签对齐，也可以忽略标签，在 Series、DataFrame 计算时自动与数据对齐；
- 强大、灵活的分组（group by）功能：拆分-应用-组合数据集，聚合、转换数据；
- 把 Python 和 NumPy 数据结构里不规则、不同索引的数据轻松地转换为 DataFrame 对象；
- 基于智能标签，对大型数据集进行切片、花式索引、子集分解等操作；
- 直观地合并（merge）、连接（join）数据集；
- 灵活地重塑（reshape）、透视（pivot）数据集；
- 轴支持结构化标签：一个刻度支持多个标签；
- 成熟的 IO 工具：读取文本文件（CSV 等支持分隔符的文件）、Excel 文件、数据库等来源的数据，利用超快的 HDF5 格式保存 / 加载数据；
- 时间序列：支持日期范围生成、频率转换、移动窗口统计、移动窗口线性回归、日期位移等时间序列功能。



<https://github.com/lcdse7en/pandas>

Basic Operations

Pandas read_file and DataFrame to_file.

pd.read_csv

- **filepath_or_buffer**: string
- **sep**: character, default ",",
- **usecols**: use name of columns
- **encoding**: string, "utf-8", "GBK"

df.to_excel

```
1 df.to_excel(  
2     excel_writer = "test.xlsx",  
3     sheet_name = "test",  
4     index = False,  
5     freeze_panes = (1,1)  
6 )
```

Get DataFrame the number of rows and columns.

```
1 # get rows  
2 len(df)  
3 df.shape[0]  
4 # get columns  
5 df.shape[1]  
6 len(df.columns)
```

DataFrame insert column.

df.insert

- **loc**: int
- **column**: string
- **value**: int, Series or array-like
- **allow_duplicates**: bool, default False

```
1 # method one  
2 df.insert(  
3     loc=0,  
4     column="ID",  
5     value=range(1, len(df) + 1)  
6 )  
7 # method two  
8 df["ID"] = range(1, len(df) + 1)
```

DataFrame sort.

df.sort_values

- **by**: string or list of string
- **ascending**: boolean or list of boolean, False: descending, True: ascending
- **inplace**: boolean

Advanced Operations