cloudera

Python 数据分析

Cloudera 大数据培训基地

重庆翰海睿智大数据科技有限公司

高级数据结构和操作类库 Pandas 进阶

课程章节

高级数据结构和操作类库 Pandas

进阶

高级数据结构和操作类库 Pandas 进阶

在本章中，您将了解到：

1. Pandas 读写文本格式的数据
2. Pandas 读写二进制数据格式
3. Pandas 使用数据库
4. Pandas 合并数据集
5. Pandas 重塑和轴向旋转
6. Pandas 数据转换

本章主题

高级数据结构和操作类库 Pandas 进阶

- Pandas 读写文本格式的数据
- Pandas 读写二进制数据格式
- Pandas 使用数据库
- Pandas 合并数据集
- Pandas 重塑和轴向旋转
- Pandas 数据转换
- 结论

读写文本格式数据 (1)

pandas 提供了一些用于将表格型数据读取为 DataFrame 对象的函数。其中 read_csv 和 read_table 会是今后用得最多的。

pandas 中的解析函数

函数	说明
read_csv	从文件、URL、文件型对象中加载带分隔符的数据。默认分隔符为逗号
read_table	从文件、URL、文件型对象中加载带分隔符的数据。默认分隔符为制表符（“\t”）
read_fwf	读取定宽列格式数据（也就是说，没有分隔符）
read_clipboard	读取剪贴板中的数据，可以看做read_table的剪贴板版。在将网页转换为表格时很有用

图 1: ‘解析函数’

读写文本格式数据 (2)

pandas 解析函数在将文本数据转换为 DataFrame 时所用到的技术分为以下几大类。

- 索引
- 类型推断和数据转换
- 日期解析
- 迭代
- 不规则数据问题

读写文本格式数据 (3)

read_csv/read_table 函数的参数

参数	说明
path	表示文件系统位置、URL、文件型对象的字符串
sep或delimiter	用于对行中各字段进行拆分的字符序列或正则表达式
header	用作列名的行号。默认为0（第一行），如果没有header行就应该设置为None
index_col	用作行索引的列编号或列名。可以是单个名称/数字或由多个名称/数字组成的列表（层次化索引）
names	用于结果的列名列表，结合header=None
skiprows	需要忽略的行数（从文件开始处算起），或需要跳过的行号列表（从0开始）
na_values	一组用于替换NA的值
comment	用于将注释信息从行尾拆分出去的字符（一个或多个）
parse_dates	尝试将数据解析为日期，默认为False。如果为True，则尝试解析所有列。此外，还可以指定需要解析的一组列号或列名。如果列表的元素为列表或元组，就会将多个列组合到一起再进行日期解析工作（例如，日期/时间分别位于两个列中）
keep_date_col	如果连接多列解析日期，则保持参与连接的列。默认为False。
converters	由列号/列名跟函数之间的映射关系组成的字典。例如，{'foo': f}会对foo列的所有值应用函数f

图 3: '函数的参数'

读写文本格式数据 (4)

read_csv/read_table 函数的参数

参数	说明
verbose	打印各种解析器输出信息，比如“非数值列中缺失值的数量”等
encoding	用于unicode的文本编码格式。例如，“utf-8”表示用UTF-8编码的文本
squeeze	如果数据经解析后仅含一列，则返回Series
thousands	千分位分隔符，如“,”或“.”
dayfirst	当解析有歧义的日期时，将其看做国际格式（例如，7/6/2012 → June 7, 2012）。默认为False
date_parser	用于解析日期的函数
nrows	需要读取的行数（从文件开始处算起）
iterator	返回一个TextParser以便逐块读取文件
chunksize	文件块的大小（用于迭代）
skip_footer	需要忽略的行数（从文件末尾处算起）

图 3: ‘函数的参数’

逐块读取文本文件

在处理很大的文件时，或找出大文件中的参数集以便于后续处理是，我们可能只想读取文件的小部分或逐块对文件进行迭代。

- nrows 参数：用于读取指定行数的数据
- chunksize 参数：用于逐块读取文件数据

将数据写出到文本格式

数据也可以被输出为分隔符格式的文本。

- DataFrame 的 to_csv 方法将数据写入文件
- 将缺失值设置为指定的标记量
- 禁止行和列标签被写入
- 写出指定的部分列数据并且指定排序
- Series 对象的 to_csv 方法将数据写入文件
- 使用更为简单 CSV 文件读取

手工处理分隔符格式

大部分存储在磁盘上的表格型数据都能用 pandas.read_table 进行加载。然而，有时还需要做一些做一些手工处理。由于接收到含有畸形行的文件而使 read_table 出毛病的情况并不少见。

CSV 语支选项

参数	说明
delimiter	用于分隔字段的单字符字符串。默认为“,”
lineterminator	用于写操作的行结束符，默认为“\r\n”。读操作将忽略此选项，它能认出跨平台的行结束符
quotechar	用于带有特殊字符（如分隔符）的字段的引用符号。默认为“"”
quoting	引用约定。可选值包括csv.QUOTE_ALL（引用所有字段）、csv.QUOTE_MINIMAL（只引用带有诸如分隔符之类特殊字符的字段）、csv.QUOTE_NONNUMERIC以及csv.QUOTE_NON（不引用）。完整信息请参考Python的文档。默认为QUOTE_MINIMAL
skipinitialspace	忽略分隔符后面的空白符。默认为False
doublequote	如何处理字段内的引用符号。如果为True，则双写。完整信息及行为请参见在线文档
escapechar	用于对分隔符进行转义的字符串（如果quoting被设置为csv.QUOTE_NONE的话）。默认禁用

图 4: ‘语支选项’

JSON 数据

JSON(JavaScript Object Notation 的简称) 已经成为通过 HTTP 请求在 Web 浏览器和其他应用程序之间发送数据的标准格式之一。

JSON 非常接近于有效的 Python 代码。许多 Python 库都可以读写 JSON 数据。

本章主题

高级数据结构和操作类库 Pandas 进阶

- Pandas 读写文本格式的数据
- Pandas 读写二进制数据格式
- Pandas 使用数据库
- Pandas 合并数据集
- Pandas 重塑和轴向旋转
- Pandas 数据转换
- 结论

pickle 序列化

实现数据的二进制格式存储和读取最简单的方法使用 pandas 对象将数据以 pickle 存储和读取。

- to_pickle 方法：用于将数据存储到二进制文件中。
- read_pickle 方法：用于读取二进制文件中的数据。

读写 Microsoft Excel 文件

python 的第三方库可以支持 Excel 文件的读写表格型数据。xlrd 库用于读取 Excel 数据、xlwt 库用于向 Excel 进行写入数据。

通过 *pip* 命令分别安装 *xlrd* 和 *xlwt* 库。

那么 Pandas 的 ExcelFile 类型支持读取存储在 Excel 中的表格型数据。

提示： xlwt 库写入文件时，存储为 xlsx 格式文件无法使用。这是它本身的 BUG。

本章主题

高级数据结构和操作类库 Pandas 进阶

- Pandas 读写文本格式的数据
- Pandas 读写二进制数据格式
- Pandas 使用数据库
- Pandas 合并数据集
- Pandas 重塑和轴向旋转
- Pandas 数据转换
- 结论

加载 MySQL

将数据从 MySQL 加载到 DataFrame 的过程很简单。

下载地址：http://www.lfd.uci.edu/~gohlke/pythonlibs/#mysql-python, 将文件下载至 D 盘 DataAnalysisPackage 文件夹中

安装 mysqlclient-1.3.12-cp36-cp36m-win_amd64.whl

本章主题

高级数据结构和操作类库 Pandas 进阶

- Pandas 读写文本格式的数据
- Pandas 读写二进制数据格式
- Pandas 使用数据库
- Pandas 合并数据集
- Pandas 重塑和轴向旋转
- Pandas 数据转换
- 结论

合并数据集介绍

Pandas 对象中的数据可以通过一些内容中的方法进行合并。

- pandas.mege 可根据一个或多个键将不同的 DataFrame 中的行连接起来。
- pandas。concat 可以沿着一个轴将多个对象堆叠到一起。
- 实例方法 combine_first 可以将重复数据编接在一起，用一个对象中的值填充另一个对象中的缺失值。

数据库风格的 DataFrame 合并 (1)

数据集的合并 (merge) 或连接 (join) 运算是通过一个或多个键将行链接起来的。这些运算是关系型数据库的核心。pandas 的 merge 函数是对数据应用这些算法的主要切入点。

参数	说明
left	参与合并的左侧DataFrame
right	参与合并的右侧DataFrame
how	“inner”、“outer”、“left”、“right”其中之一。默认为“inner”

图 5: ‘merge 函数的参数’

数据库风格的 DataFrame 合并 (2)

参数	说明
on	用于连接的列名。必须存在于左右两个DataFrame对象中。如果未指定，且其他连接键也未指定，则以left和right列名的交集作为连接键
left_on	左侧DataFrame中用作连接键的列
right_on	右侧DataFrame中用作连接键的列
left_index	将左侧的行索引用作其连接键
right_index	类似于left_index
sort	根据连接键对合并后的数据进行排序，默认为True。有时在处理大数据集时，禁用该选项可获得更好的性能
suffixes	字符串值元组，用于追加到重叠列名的末尾，默认为('_x', '_y')。例如，如果左右两个DataFrame对象都有“data”，则结果中就会出现“data_x”和“data_y”
copy	设置为False，可以在某些特殊情况下避免将数据复制到结果数据结构中。默认总是复制

图 6: ‘merge 函数的参数’

索引上的合并

有时候，DataFrame 中的连接键位于其索引中。在这种情况下，可以传入 left_index=True 或 right_index(或者两个都传) 以说明索引应该被用作连接键。

- 连接键交集
- 连接键并集
- 层次化索引的数据处理
- 同时合并双方的索引
- join 实例方法按索引合并

轴向连接 (1)

另一种数据合并运算也不称为连接、绑定或堆叠。NumPy 有一个用于合并原始 NumPy 数组的 concatenation 函数。

那么对于 pandas 对象 (如 Series 和 DataFrame)，带有标签的轴能够让我们进一步推广数组的连接运算。

pandas 的 concat 函数提供一种能够解决这些问题的可靠方式。

轴向连接 (2)

参数	说明
objs	参与连接的pandas对象的列表或字典。唯一必需的参数
axis	指明连接的轴向，默认为0
join	“inner”、“outer”其中之一，默认为“outer”。指明其他轴向上的索引是按交集（inner）还是并集（outer）进行合并
join_axes	指明用于其他n-1条轴的索引，不执行并集/交集运算
keys	与连接对象有关的值，用于形成连接轴向上的层次化索引。可以是任意值的列表或数组、元组数组、数组列表（如果将levels设置成多级数组的话）
levels	指定用作层次化索引各级别上的索引，如果设置了keys的话[译注3]
names	用于创建分层级别的名称，如果设置了keys和（或）levels的话
verify_integrity	检查结果对象新轴上的重复情况，如果发现则引发异常。默认（False）允许重复
ignore_index	不保留连接轴上的索引，产生一组新索引range(total_length)

图 7: ‘concat 函数的参数’

合并重叠数据

还有一种数据组合问题不能用简单的合并或连接运算来处理。比如，可能有索引全部或部分重叠的两个数据集。

本章主题

高级数据结构和操作类库 Pandas 进阶

- Pandas 读写文本格式的数据
- Pandas 读写二进制数据格式
- Pandas 使用数据库
- Pandas 合并数据集
- Pandas 重塑和轴向旋转
- Pandas 数据转换
- 结论

重塑层次化索引

有许多用于重新排列表格型数据的基础运算。这些函数也称为重塑或轴向旋转运算。

层次化索引为 DataFrame 数据的重排任务提供了一种具有良好一致性的方式。主要功能分为两种。

- stack: 将数据的列 “旋转” 为行
- unstack: 将数据的行 “旋转” 为列

将“长格式”旋转为“宽格式”

时间序列数据通常是以所谓的“长格式”或“堆叠格式”存储在数据库和 CSV 中的。

本章主题

高级数据结构和操作类库 Pandas 进阶

- Pandas 读写文本格式的数据
- Pandas 读写二进制数据格式
- Pandas 使用数据库
- Pandas 合并数据集
- Pandas 重塑和轴向旋转
- Pandas 数据转换
- 结论

移除重复数据

数据集操作另外一类则是过滤、清理以及其他的转换工作。

在 DataFrame 中常常会出现重复行数据。

- duplicated 方法: 返回一个布尔型 Series，表示各行是否是重复行
- drop_duplicates 方法: 返回一个移除了重复行的 DataFrame。

利用函数或映射进行数据转换

在对数据集进行转换时，可能希望根据数组、Series 或 DataFrame 列中的值来实现转换工作。

Series 的 map 方法可以接受一个函数或含有映射关系的字典型对象。

替换值

填充缺失数据可以看做值替换的一种特殊情况。虽然前面提到的 map 可以用于修改对象的数据子集，而 replace 则提供了一种实现该功能更简单和灵活的方法。

重命名轴索引

跟 Series 中的值一样，轴标签也可以通过函数或映射进行转换，从而得到一个新对象。轴还可以被修改，而无需新建一个数据结构。

离散化和面元划分

为了便于分析，连续数据常常被离散化或拆分为“面元”。

- cut 函数：用于根据范围进行面元划分
- qcut 函数：用于分位数对数据进行面元划分

检测和过滤异常值

异常值的过滤或变换运算在很大程度上其实就是数组运算。

排列和随机采样

利用 numpy.random.permutation 函数可以轻松实现对 Series 或 DataFrame 的列的排列功能。

计算指标/哑变量

另一种常用于统计建模或机器学习的转换方式是：将分类变量转换为“哑变量矩阵”或“指标矩阵”。

本章主题

高级数据结构和操作类库 Pandas 进阶

- Pandas 读写文本格式的数据
- Pandas 读写二进制数据格式
- Pandas 使用数据库
- Pandas 合并数据集
- Pandas 重塑和轴向旋转
- Pandas 数据转换
- 结论

基本要点

- Pandas 读写文本格式的数据
- 读写文本格式数据、逐块读取文本文件
- 将数据写出到文本格式、手工处理分隔符格式、JOSN 数据
- Pandas 读写二进制数据格式
- pickle 序列化、读取 Microsoft Excel 文件
- Pandas 使用数据库
- 加载 MySQL
- Pandas 合并数据集
- 数据库风格的 DataFrame 合并、索引上的合并、轴向连接、合并重叠数据
- Pandas 重塑和轴向旋转
- 重塑层次化索引、将 “长格式” 旋转为 “宽格式”
- Pandas 数据转换