pandas入门培训

讲师：林应
微博：http://weibo.com/u/2607195824
最后更新：2016/09/03

pandas简介

- 官网链接：http://pandas.pydata.org/
- pandas = pannel data + data analysis
- 简介：Pandas是python的一个数据分析包，最初由AQR Capital Management于2008年4月开发，并于2009年底开源出来，目前由专注于Python数据包开发的PyData开发team继续开发和维护，属于PyData项目的一部分。Pandas最初被作为金融数据分析工具而开发出来，因此，pandas为时间序列分析提供了很好的支持。
- 作者介绍：Wes McKinney

基本功能

- 开发pandas时提出的需求
  - 具备按轴自动或显式数据对齐功能的数据结构
  - 集成时间序列功能
  - 既能处理时间序列数据也能处理非时间序列数据的数据结构
  - 数学运算和约简（比如对某个轴求和）可以根据不同的元数据（轴编号）执行
  - 灵活处理缺失数据
  - 合并及其他出现在常见数据库（例如基于SQL的）中的关系型运算

数据结构 Series

- Series是一种类似于一维数组的对象，它由一组数据（各种NumPy数据类型）以及一组与之相关的数据标签（即索引）组成。
- Series的字符串表现形式为：索引在左边，值在右边。
- 创建
- 读写
- 运算
- 例子代码分析：introduction_to_pandas_data_structures/series.py

数据结构 DataFrame

- DataFrame是一个表格型的数据结构，它含有一组有序的列，每列可以是不同的值类型（数值、字符串、布尔值等）。
- DataFrame既有行索引也有列索引，它可以被看做由Series组成的字典（共用同一个索引）。

数据结构 DataFrame

- 可以输入给DataFrame构造器的数据

类型	说明
二维ndarray	数据矩阵，还可以传入行标和列标。
由数组、列表或元组组成的字典	每个序列会变成DataFrame的一列，所有序列的长度必须相同。
NumPy的结构化/记录数组	类似于“由数组组成的字典”
由Series组成的字典	每个Series会组成一列。如果没有显示指定索引，则各Series的索引会被合并成结果的行索引。
由字典组成的字典	各内层字典会成为一列。键会被合并成结果的行索引，跟“由Series组成的字典”的情况一样。
字典或Series的列表	各项将会成为DataFrame的一行。字典键或Series索引的并集将会成为DataFrame的列标。
由列表或元组组成的列表	类似于“二维ndarray”
另一个DataFrame	该DataFrame的索引将会被沿用，除非显示指定了其他索引。
NumPy的MaskedArray	类似于“二维ndarray”的情况，只是掩码值在结果DataFrame会变成NA/缺失值。

数据结构 DataFrame

- 创建
- 读写
- 例子代码：introduction_to_pandas_data_structures/dataframe.py

数据结构 索引对象

- pandas的索引对象负责管理轴标签和其他元数据（比如轴名称等）。构建Series或DataFrame时，所用到的任何数组或其他序列的标签都会被转换成一个Index。
- Index对象是不可修改的（immutable），因此用户不能对其进行修改。不可修改性非常重要，因为这样才能使Index对象在多个数据结构之间安全共享。
- 例子代码：introduction_to_pandas_data_structures/index_objects.py

数据结构 索引对象

- pandas中主要的index对象

类型	说明
index	最泛化的Index对象，将轴标签为一个由Python对象组成的NumPy数组。
Int64Index	针对整数的特殊Index
MultiIndex	“层次化”索引对象，表示单个轴上的多层索引。可以看做由园数组组成的数组。
DatetimeIndex	存储纳秒级时间戳
PeriodIndex	针对Period数据的特殊Index

数据结构 索引对象

- Index的方法和属性 I

类型	说明
append	append 连接另一个Index对象，产生一个新的Index。
diff	计算差集，并得到一个Index。
intersection	计算交集
union	计算并集
isin	计算一个指示各值是否包含在参数集合中的布尔型数组
delete	删除索引i处的元素，并得到新的Index。

数据结构 索引对象

- Index的方法和属性 II

类型	说明
drop	删除传入的值，并得到新的索引。
insert	将元素插入到索引i处，并得到新的Index。
is_monotonic	当各元素均大于等于前一个元素时，返回True。
is_unique	当Index没有重复值时，返回True。
unique	计算Index中唯一值得数组

基本功能 重新索引

- 创建一个适应新索引的新对象，该Series的reindex将会根据新索引进行重排。如果某个索引值当前不存在，就引入缺失值
- 对于时间序列这样的有序数据，重新索引时可能需要做一些插值处理。method选项即可达到此目的。
- 例子代码：essential_functionality/reindexing.py

基本功能 重新索引

- reindex函数的参数

类型	说明
index	用作索引的新序列。既可以是Index实例，也可以是其它序列型的Python数据结构。Index会被完全使用，就像没有任何复制一样。
method	插值填充方式，ffill或bfill。
fill_value	在重新索引过程中，需要引入缺失值时使用的替代值。
limit	前向或后向填充时的最大填充量
level	在MultiIndex的指定级别上匹配简单索引，否则选取其子集。
copy	默认为True，无论如何都复制。如果为False，则新旧相等就不复制。

基本功能 丢弃指定轴上的项

- 丢弃某条轴上的一个或多个项很简单，只要有一个索引数组或列表即可。由于需要执行一些数据整理和集合逻辑，所以drop方法返回的是一个在指定轴上删除了指定值的新对象
- 例子代码：essential_functionality/dropping_entries_from_an_axis.py

基本功能 索引、选取和过滤

- Series索引（obj[...]）的工作方式类似于NumPy数组的索引，只不过Series的索引值不只是整数。
- 利用标签的切片运算与普通的Python切片运算不同，其末端是包含的（inclusive）。
- 对DataFrame进行索引其实就是获取一个或多个列
- 为了在DataFrame的行上进行标签索引，引入了专门的索引字段ix。
- 例子代码：essential_functionality/indexing_selection_and_filtering.py

基本功能 索引、选取和过滤

- DataFrame的索引选项

类型	说明
obj[val]	选取DataFrame的单个列或一组列。在一些特殊情况下会比较便利：布尔型数组（过滤行）、切片（行切片）、布尔型DataFrame（根据条件设置值）。
obj.ix[val]	选取DataFrame的单个行或一组行
obj.ix[:, val]	选取单个列或列子集
obj.ix[val1, val]	同时选取行或列
reindex方法	将一个或多个轴匹配到新索引
xs方法	根据标签选取单行或单列，并返回一个Series。
icol、irow方法	根据整数位置选取单行或单列，并返回一个Series。
get_value、set_value方法	根据行标签或列标签选取单个值

基本功能 算术运算和数据对齐

- 对不同的索引对象进行算术运算
- 自动数据对齐在不重叠的索引处引入了NA值，缺失值会在算术运算过程中传播。
- 对于DataFrame，对齐操作会同时发生在行和列上。
- fill_value参数
- DataFrame和Series之间的运算
- 例子代码：essential_functionality/arithmetic_and_data_alignment.py

基本功能 函数应用和映射

- numpy的ufuncs（元素级数组方法）
- DataFrame的apply方法
- 对象的applymap方法（因为Series有一个应用于元素级的map方法）
- 例子代码：essential_functionality/function_application_and_mapping.py

基本功能 排序和排名

- 对行或列索引进行排序
- 对于DataFrame，根据任意一个轴上的索引进行排序
- 可以指定升序降序
- 按值排序
- 对于DataFrame，可以指定按值排序的列
- rank函数
- 例子代码：essential_functionality/sorting_and_ranking.py

基本功能 带有重复值的索引

- 对于重复索引，返回Series，对应单个值的索引则返回标量。
- 例子代码：essential_functionality/axis_indexes_with_duplicate_values.py

汇总和计算描述统计

- 常用方法选项

类型	说明
axis	指定轴，DataFrame的行用0，列用1。
skipna	排除缺失值，默认值为True。
level	如果轴是层次化索引的（即MultiIndex），则根据level选取分组。

汇总和计算描述统计

- 常用描述和汇总统计函数 I

类型	说明
count	非NA值的数量
describe	针对Series或各DataFrame列计算汇总统计
min, max	计算最小值和最大值
argmin, argmax	计算能够获取到最小值和最大值的索引位置（整数）
idxmin, idxmax	计算能够获取到最小值和最大值的索引值
sum	值的总和
mean	值的平均数
median	值的算术中位数
mad	根据平均值计算平均绝对离差

汇总和计算描述统计

- 常用描述和汇总统计函数 II

类型	说明
var	样本值的方差
std	样本值的标准差
skew	样本值的偏度（三阶矩）
kurt	样本值的偏度（四阶矩）
cumsum	样本值的累计和
cummin, cummax	样本值的累计最大值和累计最小值
cumprod	样本值的累计积
diff	计算一阶差分
pct_change	计算百分数变化

汇总和计算描述统计

- 数值型和非数值型的区别
- NA值被自动排查，除非通过skipna选项
- 例子代码：summarizing_and_computing_descriptive_statistics/intro.py

汇总和计算描述统计 相关系数与协方差

- 相关系数：相关系数是用以反映变量之间相关关系密切程度的统计指标。百度百科
- 协方差：从直观上来看，协方差表示的是两个变量总体误差的期望。如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值时另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值；如果两个变量的变化趋势相反，即其中一个变量大于自身的期望值时另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。
- 例子代码分析：summarizing_and_computing_descriptive_statistics/correlation_and_covariance.py

汇总和计算描述统计 唯一值以及成员资格

- 常用方法

类型	说明
is_in	计算一个表示“Series各值是否包含于传入的值序列中”的布尔型数组
unique	计算Series中的唯一值数组，按发现的顺序返回。
value_counts	返回一个Series，其索引为唯一值，其值为频率，按计数值降序排列。

- 例子代码：
  correlation_and_covariance/unique_values_value_counts_and_membership.py

处理缺失数据

- NA处理方法

类型	说明
dropna	根据各标签的值中是否存在缺少数据对轴
fillba	样本值的标准差
isnull	样本值的偏度（三阶矩）
notnull	

- NaN（Not a Number）表示浮点数和非浮点数组中的缺失数据
- None也被当作NA处理
- 例子代码分析：handling_missing_data/intro.py

处理缺失数据 滤除缺失数据

- dropna
- 布尔索引
- DatFrame默认丢弃任何含有缺失值的行
- how参数控制行为，axis参数选择轴，thresh参数控制留下的数量
- 例子代码分析：handling_missing_data/filtering_out_missing_data.py

处理缺失数据 填充缺失数据

- fillna
- inplace参数控制返回新对象还是就地修改
- 例子代码分析：handling_missing_data/filling_in_missing_data

层次化索引

- 使你能在一个轴上拥有多个（两个以上）索引级别。抽象的说，它使你能以低纬度形式处理高维度数据。
- 通过stack与unstack变换DataFrame
- 例子代码分析：hierarchical_indexing/intro.py

层次化索引 重新分级顺序

- 索引交换
- 索引重新排序
- 例子代码：hierarchical_indexing/reordering_and_sorting_levels.py

层次化索引 根据级别汇总统计

- 指定索引级别和轴
- 例子代码：hierarchical_indexing/summary_statistics_by_level.py

层次化索引 使用DataFrame的列

- 将指定列变为索引
- 移除或保留对象
- reset_index恢复
- 例子代码：hierarchical_indexing/using_a_dataframes_columns.py

其它话题 整数索引

- 歧义的产生
- 可靠的，不考虑索引类型的，基于位置的索引。
- 例子代码分析：other_pandas_topics/integer_indexing.py

其它话题 面板(Pannel)数据

- 通过三维ndarray创建pannel对象
- 通过ix[...]选取需要的数据
- 访问顺序：item -> major -> minor
- 通过stack展现面板数据
- 例子代码分析：other_pandas_topics/panel_data.py