Pandas教程一：基础概念

大家好，这里是Python知识学堂，最近很多小伙伴在后台提问关于Pandas的相关问题，我发现这些问题都是关于Pandas的基础知识。虽然很多公众号都写了关于Pandas教程，大多数是关于pandas教程都是一些高级或实战炫技，但是很多都不是我们需要的基础教程，为此本公众号将开始出一点基础的教程。

1. Pandas简介

Pandas是一个开源的，BSD许可的库。其提供了快速、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的目标是为Python编程语言提供高性能，易于使用的数据结构和数据分析工具。

1. Pandas适合处理的数据类型

（1）具有异构类型列的表格数据（不同列的数据类型不相同，如有的列为str，有的为int，有的为datetime），比如我们常见的SQL表或Excel电子表格数据；

（2）有序和无序（且不一定是固定频率）的时间序列数据；

（3）具有行和列标签的任意矩阵数据（同质或者异构，不同列数据类型相同则同质）；

（4）任何其他形式的观察/统计数据集，这些数据在被存放到pandas的数据结构当中时，不需要打上标签；

1. Pandas数据结构

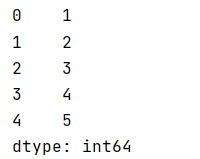
Pandas的数据结构主要有两种，Series和DataFrame，我们来看一下他们的区别：

|  |  |  |
| --- | --- | --- |
| **维度** | 数据结构名称 | **描述** |
| 1 | Series | 带标签的一维同构数组 |
| 2 | DataFrame | 带标签的，大小可变的，二维异构表格 |

我们生成一个Series和DataFrame来看看他们的形式：

import pandas as pd  
test\_list = [1,2,3,4,5]  
t\_series = pd.Series(test\_list)  
  
data = {  
 **'name'**:[**'Beijing'**,**'Shanghai'**,**'Xiamen'**,**'Wuhan'**,**'Chizhou'**],  
 **'year'**:[2000,2001,2002,2001,2002],  
 **'stage'**:[1,0,1,0,0]  
}  
t\_dataframe = pd.DataFrame(data)  
print(t\_series)  
print(t\_dataframe)

输出如下：





具体关于Series和DataFrame的介绍我们将在后期的推文中展开。

需要注意的是Pandas 所有数据结构的值都是可变的，但数据结构的大小并非都是可变的，如Series 的长度不可改变，但 DataFrame 里就可以插入列数据。

1. Pandas优势

（1）缺失数据的处理：处理浮点与非浮点数据里的缺失数据，表示为NaN；

（2）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；

（3）强大灵活的分组功能：拆分-应用-组合数据集，聚合、转换数据；

（4）基于NumPy的Pandas可将不规则、不同索引的数据轻松地转换为 DataFrame 对象；

（5）Pandas 速度很快。Pandas的很多底层算法都用优化过。

（6）基于智能标签，可对大型数据集进行切片、花式索引、子集分解等操作；

（7）其他优势等等

1. 总结

以上就是本次推文的全部内容，主要介绍了Pandas的一些基础知识，包括可处理的数据类型，自身支持的数据结构，以及在数据处理中的一些优势。相信大家已经全部掌握了，下期关于Pandas的推文，我们将开始介绍其数据结构Series的使用和常用方法。