## 法律声明

□ 本课件包括:演示文稿,示例,代码,题库,视频和声音等,小象学院拥有完全知识产权的权利;只限于善意学习者在本课程使用,不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意,我们将保留一切通过法律手段追究违反者的权利。

- □ 课程详情请咨询
  - 微信公众号: 大数据分析挖掘
  - 新浪微博: ChinaHadoop







# 零基础Python入门

--梁斌





## 空气质量指数计算 9.0



### 案例描述

- 为了能有效地提取并利用网络信息并工作提高效率,出现了网络爬虫
- 利用网络爬虫实时获取城市的空气质量
- 利用beautifulsoup4获取所有城市的空气质量
- 将获取的所有城市空气质量保存成CSV数据文件
- 利用Pandas进行数据处理分析



## 什么是Pandas

#### **Pandas**

- 一个强大的分析结构化数据的工具集
- · 基础是NumPy,提供了高性能矩阵的运算
- 应用,数据挖掘,数据分析
  - 如,学生成绩分析、股票数据分析等。
- 提供数据清洗功能





#### **Series**

- 类似一维数组的对象
- 通过list构建Series
  - ser\_obj = pd.Series(range(10))
- 由数据和索引组成
  - 索引在左,数据在右
  - 索引是自动创建的
- 获取数据和索引
  - ser\_obj.index, ser\_obj.values
- 预览数据
  - ser obj.head(n)

#### SERIES

index	element
IIIUEA	

0	1
1	2
2	3
3	4
4	5



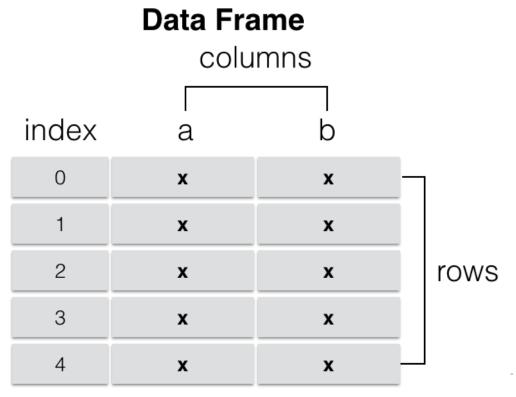
#### Series (续)

- 通过索引获取数据
  - ser\_obj[idx]
- 索引与数据的对应关系仍保持在数组运算的结果中
- 通过dict构建Series
- name属性
  - ser obj.name, ser obj.index.name



#### **DataFrame**

- 类似多维数组/表格数据 (如 , excel, R中的data.frame)
- 每列数据可以是不同的类型, what about ndarray?
- 索引包括列索引和行索引





#### **DataFrame**

- 通过ndarray构建DataFrame
- 通过dict构建DataFrame
- 通过列索引获取列数据(Series类型)
  - df\_obj[col\_idx] 或 df\_obj.col\_idx
- 增加列数据,类似dict添加key-value
  - df\_obj[new\_col\_idx] = data
- 删除列
  - del df\_obj[col\_idx]



### Pandas的数据操作

#### 索引操作

- DataFrame索引
  - 列索引
    - df\_obj[ 'label' ]
  - 不连续索引
    - df\_obj[[ 'label1' , 'label2' ]]



### Pandas的数据操作

#### 排序

- · sort\_index , 索引排序
  - 对DataFrame操作时注意轴方向
- 按值排序
  - sort\_values(by= 'label' )



## Pandas统计计算和描述

#### 常用的统计计算

- sum, mean, max, min...
- axis=0 按列统计,axis=1按行统计
- skipna 排除缺失值 , 默认为True
- idmax, idmin, cumsum

#### 统计描述

describe 产生多个统计数据



# Pandas统计计算和描述

方法	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
quantile	计算样本的分位数(0到1)
sum	值的总和
mean	值的平均数
median	值的算术中位数(50%分位数)
mad	根据平均值计算平均绝对离差
var	样本值的方差
std	样本值的标准差



# Pandas统计计算和描述

方法	说明
skew	样本值的偏度(三阶矩)
kurt	样本值的峰度(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 (对时间序列很有用)
pct_change	计算百分数变化



### Next?

- 数据清洗
- 利用Pandas进行数据可视化



### 疑问

□问题答疑: <a href="http://www.xxwenda.com/">http://www.xxwenda.com/</a>

■可邀请老师或者其他人回答问题

小象问答邀请 @Robin\_TY 回答问题





#### 联系我们

#### 小象学院: 互联网新技术在线教育领航者

- 微信公众号: 小象

- 新浪微博: ChinaHadoop



