法律声明

口本淉件包括：渍示文稿，示例，代码，题库，视频和声 音等，小象羊院拥有完全知识产权的权利；只限于善意 争习者在本淉程使用，不得在淉程范围外向任何笫三方 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。

回•.•

口淉程详情请咨询

* 微信公众号：大数据分析挖掘
* 新浪微博： ChinaHadoop



互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn



零基础Python入门

－－梁斌

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

第九讲





AIR QUALITY INDEX I

空气质量指数计算9.0

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

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bea utifu I sou p4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

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

什么是Pandas

Pandas

* —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



互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

Pandas的数据结构

|  |  |  |
| --- | --- | --- |
| Series | SERIES | |
| • 类似—维数组的对象 | index | element |
| • 通过list构建Series | 。 | 1 |
| • ser \_ obj = pd.Series(range(10)) |  |
| • 由数据和索引组成 | 1 | 2 |
| • 索引在左，数据在右 | 2 | 3 |
| • 索引是自动创建的 | 3 | 4 |
| • 获取数据和索引 |  |  |
| • ser \_ obj.index, ser \_ obj.values | 4 | 5 |
| • 预览数据 |  |  |
| • ser \_ obj.head(n) |  |  |

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

Pandas的数据结构

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diet构建Series
* name属性

• ser \_ obj.name, ser \_ obj.index.name

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

Pandas的数据结构

Data Frame

• 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

Data Frame

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | columns |  |  |
|  |  |  |  | I |  |
| index | | a |  | b |  |
| 。 |  | X |  | X |  |
|  | 一， |  |  |  |  |
|  | ＇』 |  |  |  |  |
| 1 |  | X |  | X |  |
|  | 一， |  |  |  |  |
|  | 一， |  | . r |  |  |
| 2 |  | X |  | X |  |
| 3 |  | X |  | X |  |
| 4 |  | X |  | X |  |
|  |  |  |  |  |  |
| 互联网新技术在线教育领航者 |  |  |  |  | 匕J.I L iilL丁-r兀 |
|  |  |  |  | ChinaHadoop.cn |



Pandas的数据结构

Data Frame

* 通过ndarray构建Data Frame
* 通过diet构建Data Frame
* 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diet添加key-value

• df \_ obj [new\_ col\_idx] = data

• 删除列

• del df \_ obj [ col\_idx]

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

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']]

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

Pandas的数据操作

排序

• sort\_i ndex , 索引排序

• 对Data Frame操作时注意轴方向 • 按值排序

• sort\_ values(by='label')

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

Pandas统计计算和描述

常用的统计计算

* sum, mean, max, min…
* axis=O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

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

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Data Frame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(o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

互联网新技术在线教育领航者 E笥山蛊学陲

ChinaHadoop.cn

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

cummm、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—阶差分（对时间齐列很有用）

pct\_change 计算百分数变化

互联网新技术在线教育领航者 E笥山蛊学陲

ChinaHadoop.cn

Next?

* 数据清洗
* 利用Pandas进行数据可视化



互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

疑问

口问题答疑： <http://www.xxwenda.com/>

·可邀清老师或者其他人回答问题

小象问答邀请@Robin\_TY回答问题



互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

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

－亲Ii讳良很女博： ChinaHadoop



互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn