Python 进阶训练营

尹会生

9 pandas

目　录

CONTENTS

学习目标

1.使用 pandas 清洗爬虫收集的数据

数据集

kaggle:
https://www.kaggle.com

天池:
https://tianchi.aliyun.com/dataset

搜狗实验室:
http://www.sogou.com/labs/resource/list_pingce.php

DC竞赛:
https://www.pkbigdata.com/common/cmptIndex.html

DF竞赛:
https://www.datafountain.cn/datasets

pandas 的基本数据结构

Series 数据结构	DataFrame 数据结构
类似一维数组，但是带索引 Series 有 values 和 index 属性	DataFrame 是表格型数据，类似 excel

数据的获取和预处理

数据的导入和新建

- read_excel() 支持从 .xlsx 导入数据
- read_csv() 支持从 csv 导入数据
- read_table() 支持从有规则的 txt 导入数据

数据预处理：缺失、重复、异常值的处理方法

- 缺失值一般用 NaN 表示，可以用 dropna() 删除行
- 缺失值也可以用 fillna() 填充
- 重复值用 drop_duplicates() 删除

索引设置

- set_index() 设置索引

数据的操作

数据的排序操作

使用 sort_values() 进行排序，可以按升序和降序排列

数据行列选择与查找操作

数据行列互换操作

行列互换也称作转置，T 方法表示

数据的运算

算数运算

行列之间的加减乘除

比较运算

> < !=(不等于)

汇总与统计运算

sum()
count()
mean()
max()
median()
min()
mode()
var()
std()

数据透视表

和 excel 透视表原理相同

数据的导出

导出为 .csv 格式　　to_csv()
导出为 excel 格式　　to_excel()

THANKS!