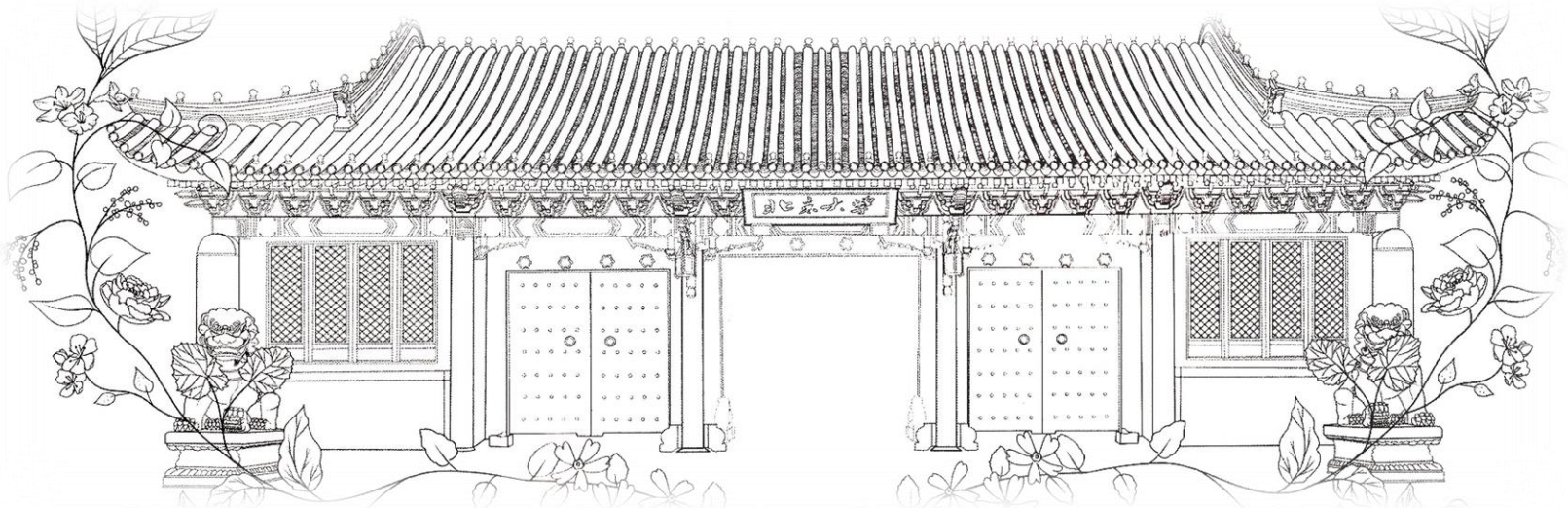
前 3 周内容回顾与提炼

北京大学信息管理系

2021.3.29.Monday



Python基本语法

- 四种数据结构（set, list, tuple, dict）
- 两种表达形式（列表推导式、生成器表达式）
- 两种函数（普通函数、匿名函数lambda）
- 模块化

numpy

- numpy简介
  - 创建与打印数组（np.array()）
  - 基本运算（+、-、两种乘法、通用函数）
  - 索引、切片和迭代
  - 数组的形状操作、分割和组合
  - 复制和视图（注意和前面提到的赋值、拷贝的区别与联系！）

pandas

- 基本数据结构的生成
  - Series：传list或dict进去
  - DataFrame：dict套list、dict套dict
- 索引是不可变的，但可以重新索引reindex()，填补方式有ffill()和mfill()
- 从坐标轴删除条目：drop()，注意有inplace()

pandas

- 索引、选择和过滤
  - Series：用整数下标索引切片、标签切片(包含end)索引
  - DataFrame：
    - df[‘列名’], df[[‘列名1’,’列名2’]]
    - df[行整数下标或下标切片]
    - 通过轴标签：df.loc[label], df.loc[:, label], df.loc[label_1, label_2]
    - 通过整数下标：df.iloc[where], df.iloc[:, where], df.iloc[where _1, where _2]

pandas

- 算术和数据对齐
  - NA值会传播
  - add, sub, div… (fill_value可以设置)
- DataFrame和Series之间的操作
  - 默认地：DataFrame和Series间的算术运算Series的索引将匹配DataFrame的列，并在行上扩展
- 函数应用和映射：
  - Series：map方法
  - DataFrame：apply用于某一行或列、applymap用于每一个元素

pandas

- 排序
  - sort_index() vs. sort_values()，默认升序
- 排名
  - rank()方法，可以指定行或列

pandas

- 描述性统计
  - 汇总：count(), describe(), max()…
  - 唯一值：unique(), value_counts()…
  - 成员判断：isin()

pandas

- 缺失值处理
  - dropna()：默认剔除所有包含缺失值的行
    - 参数可调：how='all'只剔除全部NAN的行，thresh设置阈值
  - fillna()：可对每列填充不同值，可指定原地修改
    - 注意inplace, axis, method, value等参数
  - isnull(), notnull()

pandas

- 数据读写
  - read()简单顺序读取, readline()每次读取一行, readlines()按行读取所有内容
  - 使用pandas读文件
    - 表头、跳几行、缺失值…
  - 使用pandas写文件
    - to_csv, to_json, …

pandas

- 数据清洗
  - 去重
  - 利用函数或映射进行转换
- 数据聚合和分组
  - Groupby()：最简单的分组法、使用字典或Series分组、使用函数分组、使用索引级别分组

前3周的练习

- Python语法练习
- Numpy练习
- Pandas练习A、B、C、D