Python 中级培训简要

数据分析&可视化入门，此系草稿，时间仓促可能有错误，请大家包容并提出建议随时沟通。张博华于山东·威海·荣成 19.11.1

前置部分理论知识

itchat

https://itchat.readthedocs.io/zh/latest/
itchat 是一个开源的微信个人号接口，使用 python 调用微信从未如此简单。
使用不到三十行的代码，你就可以完成一个能够处理所有信息的微信机器人。
当然，该 api 的使用远不止一个机器人，更多的功能等着你来发现，比如这些。
该接口与公众号接口 itchatmp 共享类似的操作方式，学习一次掌握两个工具。

NumPy

https://numpy.org/， https://www.numpy.org.cn/
NumPy 是使用 Python 进行科学计算的基础包。它包含如下的内容:
一个强大的 N 维数组对象。
复杂的（广播）功能。
用于集成 C / C ++和 Fortran 代码的工具。
有用的线性代数，傅里叶变换和随机数功能。
除了明显的科学用途外，NumPy 还可以用作通用数据的高效多维容器。可以定义任意数据类型。这使 NumPy 能够无缝快速地与各种数据库集成。
NumPy 是在 BSD 许可下获得许可的，允许重用而不受限制。

Pandas

https://pandas.pydata.org/， https://www.pypandas.cn/
Pandas 是一个开源的，BSD 许可的库，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。
Pandas 是 NumFOCUS 赞助的项目。这将有助于确保 Pandas 成为世界级开源项目的成功，并有可能捐赠给该项目。
Pandas 是基于 NumPy 的一种工具，该工具是为了解决数据分析任务而创建的。Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具。Pandas 提供了大量能使我们快速便捷地处理数据的函数和方法。你很快就会发现，它是使 Python 成为强大而高效的数据分析环境的重要因素之一。

Matplotlib

https://matplotlib.org/,
Matplotlib 是一个 Python 2D 绘图库，它以多种硬拷贝格式和跨平台的交互式环境生成出版物质量的图形。Matplotlib 可用于 Python 脚本，Python 和 IPython Shell，Jupyter 笔记本，Web 应用程序服务器和四个图形用户界面工具包。
Matplotlib 尝试使容易的事情变得容易，使困难的事情变得可能。您只需几行代码就可以生成图表，直方图，功率谱，条形图，误差图，散点图等。有关示例，参见示例图和缩略图库。为了简单绘图，该 pyplot 模块提供了类似于 MATLAB 的界面，尤其是与 IPython 结合使用时。对于高级用户，您可以通过面向对象的界面或 MATLAB 用户熟悉的一组功能来完全控制线型，字体属性，轴属性等。

Seaborn

Seaborn 是基于 matplotlib 的图形可视化 python 包。它提供了一种高度交互式界面，便于用户能够做出各种有吸引力的统计图表。
Seaborn 是在 matplotlib 的基础上进行了更高级的 API 封装，从而使得作图更加容易，在大多数情况下使用 seaborn 能做出很具有吸引力的图，而使用 matplotlib 就能制作具有更多特色的图。应该把 Seaborn 视为 matplotlib 的补充，而不是替代物。同时它能高度兼容 numpy 与 pandas 数据结构以及 scipy 与 statsmodels 等统计模式。

Pillow（PIL）

https://pillow.readthedocs.io/en/stable/
PIL：Python Imaging Library，已经是 Python 平台事实上的图像处理标准库了。PIL 功能非常强大，但 API 却非常简单易用。
由于 PIL 仅支持到 Python 2.7，加上年久失修，于是一群志愿者在 PIL 的基础上创建了兼容的版本，名字叫 Pillow，支持最新 Python 3.x，又加入了许多新特性，因此，我们可以直接安装使用 Pillow。

wordcloud

http://amueller.github.io/word_cloud/
词云图，也叫文字云，是对文本中出现频率较高的“关键词”予以视觉化的展现，词云图过滤掉大量的低频低质的文本信息，使得浏览者只要一眼扫过文本就可领略文本的主旨。
安装注意事项：
要先安装微软 visual c++ 14 或以上版本全部支持包。Python 有些包安装时候需要编译，在 windows 下编译可能就要用到微软的东西，linux 下编译就用的 gnu c++的东西。

wordcloud 取决于 NumPy 和 Pillow。
要将 wordcloud 保存到文件中，matplotlib 也可以安装。请参阅下面的示例。
如果您的 python 版本没有可用的轮子，则安装软件包需要安装 C 编译器。在安装编译器之前，报告一个描述了所使用的 python 版本和操作系统的问题。
Licensing
The wordcloud library is MIT licenced, but contains DroidSansMono.ttf, a true type font by Google, that is apache licensed. The font is by no means integral, and any other font can be used by setting the font_path variable when creating a WordCloud object.

Jieba

https://github.com/fxsjy/jieba
“结巴”中文分词：做最好的 Python 中文分词组件。它主要有以下 3 种特性：
支持 3 种分词模式：精确模式、全模式、搜索引擎模式。支持繁体分词，支持自定义词典。

openpyxl

https://openpyxl.readthedocs.io/en/stable/
openpyxl 是一个 Python 库，用于读取/写入 Excel 2010 xlsx / xlsm / xltx / xltm 文件。
它的诞生是因为缺少可从 Python 本地读取/写入 Office Open XML 格式的库。

xlrd

https://github.com/python-excel/xlrd
在任何平台上都可以从 Excel 电子表格（.xls 和.xlsx，2.0 版及更高版本）中提取数据。纯 Python（2.7，3.4+）。对 Excel 日期的强大支持。Unicode 感知。该程序包最初是由 David Giffin 开发的名为“ xlreader”的实用程序的一部分从 C 到 Python 的翻译开始的。

re

https://docs.python.org/zh-cn/3/library/re.html，这是个标准库，上面的全是第三方库。
正则表达式是一个特殊的字符序列，能帮助你方便的检查一个字符串是否与某种模式匹配。
Python 自 1.5 版本起增加了 re 模块，它提供 Perl 风格的正则表达式模式。
re 模块使 Python 语言拥有全部的正则表达式功能。
compile 函数根据一个模式字符串和可选的标志参数生成一个正则表达式对象。该对象拥有一系列方法用于正则表达式匹配和替换。re 模块也提供了与这些方法功能完全一致的函数，这些函数使用一个模式字符串做为它们的第一个参数。

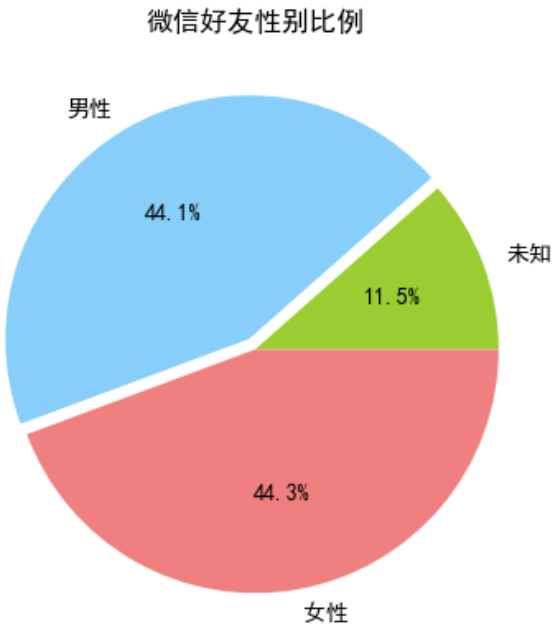
课程内容框架

第一小节：

```
# encoding:utf-8
# 抓取微信好友信息，保存到本地。
import itchat
from pandas.core.frame import DataFrame

# 形成一个二维码，微信扫描后，模拟登录网页版微信。传入 hotReload=True，生成一个静态文件 itchat.pkl ，用于存储登陆的状态，不有每次运行都登录。
itchat.auto_login(hotReload=True)
# 返回完整的好友列表，每个好友为一个字典，其中第一项为本人的账号信息。传入 update=True，将更新好友列表并返回。
friends = itchat.get_friends(update=True)[1:]  # [1:]表示切片，除掉第一项也就是本人的账号信息。

# DataFrame 是 pandas 库的一个类，我们通过调用 DataFrame()创建一个对象实例。是一种类似于 excel 的数据结构，是一种二维表，单元格可以存放数值、字符串等。
data = DataFrame(friends)

'''
to_csv()是 DataFrame 类的方法.
CSV 格式比 Excel 格式具备的优势:
1）CSV 是纯文本文件，支持追加模式写入，节省内存。Excel 是结构复杂的二进制文件，只支持一次性写入，较费内存。
2）CSV 的文件行数没有限制，在实际项目中我们已输出过上千万行的 CSV 文件；32 位系统下 Excel 单个 Sheet 最多支持 65535 行。
3）CSV 是纯文本文件，可以使用任何文本编辑器进行编辑，因此可以在 Linux 终端下对其进行修改。Excel 是二进制文件，目前已知的编辑工具有 Office，OpenOffice，WPS，都为 GUI 工具，不支持在终端下编辑。
'''
data.to_csv('we_chat_list.csv', encoding='utf_8_sig')
```

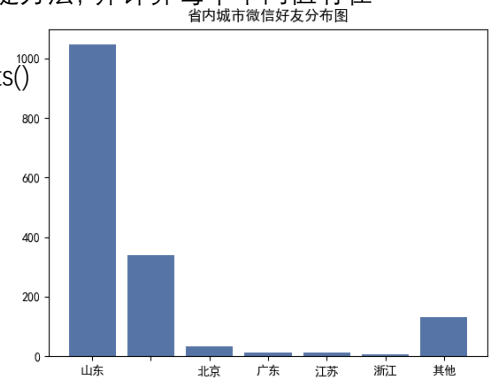
第二小节：

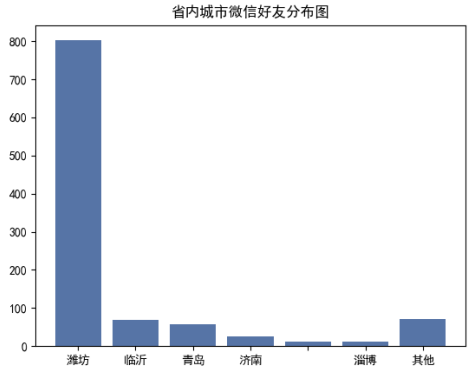
```
# encoding:utf-8
# 性别分析
import itchat
import matplotlib.pyplot as plt

itchat.auto_login(hotReload=True)
friend = itchat.get_friends(update=True)[1:]

# 遍历好友字典，对不同性别计数
sexDict = {}
total = len(friend)
for friend in friend:
    if friend['Sex'] not in sexDict:
        sexDict[friend['Sex']] = 0
    sexDict[friend['Sex']] += 1

# 不男不女性别数量赋值给变量
unkown = sexDict[0]
# 男性数量赋值给变量
male = sexDict[1]
# 女性数量赋值给变量
female = sexDict[2]

# 字体
plt.rcParams['font.sans-serif'] = ['SimHei']
# 各块颜色
colors = ['yellowgreen', 'lightskyblue', 'lightcoral']
# 饼状图外侧显示的说明文字
labels = ['未知', '男性', '女性']
# 离开中心的距离，用于突出显示其中几块
explode = (0, 0.05, 0.)
# 标题
plt.title('微信好友性别比例')
# 绘制饼图的函数
plt.pie([unkown, male, female], labels=labels, explode=explode, colors=colors,
autopct='%1.1f%%')
# 输出到文件
plt.savefig('Sex_Pie.png')
# 显示饼状图
plt.show()
```



第三小节：

```
# encoding:utf-8
# 主要分布在哪些省份（城市）
import itchat
import pandas as pd
import matplotlib.pyplot as plt  # Matplotlib 是 Python 中用的最多的 2D 图形绘图库，学好 Matplotlib 的用法可以帮助我们在统计分析中更灵活的展示各种数据的状态。
import seaborn as sns  # Seaborn 是基于 matplotlib 的图形可视化 python 包。它提供了一种高度交互式界面，便于用户能够做出各种有吸引力的统计图表。

itchat.auto_login(hotReload=True)
friends = itchat.get_friends(update=True)[1:]
data = pd.DataFrame(friends)

# 6 个默认的颜色循环主题： deep, muted, pastel, bright, dark, colorblind
# desat：每种颜色去饱和的比例
# 设置调色板
sns.set_palette('deep', desat=.8)
# 设置 rc 参数显示中文标题，设置字体为 SimHei 显示中文。 rc：run configuration。
plt.rcParams['font.sans-serif'] = ['SimHei']
# 从 Province（省份）为山东的列中将 City 对应的城市进行计数。
# value_counts()是一种查看表格某列中有多少个不同值的快捷方法，并计算每个不同值有在该列中有多少重复值。
counters = data[data['Province'] == '山东']['City'].value_counts()
# 将 Province（省份）进行计数。
# counters = data['Province'].value_counts()
# ascending 为假时降序排列。
counters = counters.sort_values(ascending=False)

length = 6  # 预设柱状图显示 6 个城市
if len(counters) > length:
    counters, temp = counters[:length], counters[length:]  # 给变量命名的另外一种方式
    value = sum(temp)  # 将 length 数量之外的城市数量求和
    counters['其他'] = value  # 在 Series(可理解为字典）中插入一组数据（键值对）
# bar 是柱状图函数，counters.index 表示 X 轴（横坐标）、counters 表示 Y 轴（纵坐标）。
plt.bar(counters.index, counters)
# 标题
plt.title('省内城市微信好友分布图')
# 输出到文件
plt.savefig('City_Bar.png')
# 显示柱状图
plt.show()
```





第四小节：


```
# encoding:utf-8
# 词云
import itchat
import numpy as np
from wordcloud import WordCloud
from wordcloud import ImageColorGenerator
from PIL import Image  #
from jieba import analyse
from pandas.core.frame import DataFrame
import re
import jieba
import matplotlib.pyplot as plt

itchat.auto_login(hotReload=True)
friends = itchat.get_friends(update=True)[1:]

dt = DataFrame(friends)
# dt.to_csv('ok.csv', encoding='utf_8_sig', header=True, index=True)

signatures = ''
for friend in friends:
    signature = friend['Signature']
    if len(signature) == 0:
        continue
    signature = signature.strip().replace('span', '').replace('class', '').replace('emoji', '')  # 去除无关数据

    signature = re.sub(r'1f(\d.+)', '', signature)
    signatures += ' '.join(jieba.analyse.extract_tags(signature))  # 关键字提取
    signatures += ' '

# Image.open，读取指定图片。

im = Image.open('hbz.jpg')  # 可替换你喜欢的图片，在当前文件夹下（相对路径）

# np.array，将读入的 im 转换成背景图数据。
mask = np.array(im)
# WordCloud 函数，建立词云对象
# mask 参数用于设置词云形状，默认的是矩形，可以读入自己选定的图片。margin：画布偏移，默认 2 像素.
word_cloud = WordCloud(font_path='simhei.ttf', background_color='white', max_words=1200, mask=mask, margin=15)
```

# generate，向 word_could 这个 WordCloud 对象中加载 signatures（文本内容），对全部文本进行自动分词（但是对中文支持不好）
word_cloud.generate(signatures)
# ImageColorGenerator 函数通过 mask 参数生成词云颜色值
image_colors = ImageColorGenerator(mask)
# 用 recolor 方法重置词云颜色为（color_func=image_colors）
word_clour = word_cloud.recolor(color_func=image_colors)
# figure 函数中，figsize 表示输出的绘图对象的宽和高、dpi 表示指定绘图对象的分辨率，即每英寸多少个像素，缺省值为 80。
plt.figure(figsize=(12, 12), dpi=100)
# imshow 函数用于对按照样本图片重置颜色的图像进行处理，并显示其格式，但是不能显示。
plt.imshow(word_clour)
# 不显示坐标尺寸
plt.axis('off')
# 显示词云图
plt.show()
# 输出到文件
word_cloud.to_file('signatures.png')

第五小节：

# encoding:utf-8
# By Zhang Bohua, from Weifang Shandong
# 将前面代码合成
import itchat
import numpy as np
from wordcloud import WordCloud
from wordcloud import ImageColorGenerator
from PIL import Image #
from jieba import analyse
import pandas as pd
import seaborn as sns # Seaborn 是基于 matplotlib 的图形可视化 python 包。它提供了一种高度交互式界面，便于用户能够做出各种有吸引力的统计图表。
import re
import jieba
import matplotlib.pyplot as plt

# 形成一个二维码，微信扫描后，模拟登录网页版微信。传入 hotReload=True，生成一个静态文件 itchat.pkl ，用于存储登陆的状态，不有每次运行都登录。
itchat.auto_login(hotReload=True)
# 返回完整的好友列表，每个好友为一个字典，其中第一项为本人的账号信息。传入

update=True，将更新好友列表并返回。
friends = itchat.get_friends(update=True)[1:] # [1:]表示切片，除掉第一项也就是本人的账号信息。
# DataFrame 是 pandas 库的一个类，我们通过调用 DataFrame()创建一个对象实例。是一种类似于 excel 的数据结构，是一种二维表，单元格可以存放数值、字符串等。
# data = DataFrame(friends)
data = pd.DataFrame(friends)
data.to_csv('we_chat_list.csv', encoding='utf_8_sig')

```
def sex_pie():
    # 遍历好友字典，对不同性别计数
    sex_dict = {}
    # total = len(friends)
    for friend in friends:
        if friend['Sex'] not in sex_dict:
            sex_dict[friend['Sex']] = 0
        sex_dict[friend['Sex']] += 1

    # 不男不女性别数量赋值给变量
    unknow = sex_dict[0]
    # 男性数量赋值给变量
    male = sex_dict[1]
    # 女性数量赋值给变量
    female = sex_dict[2]

    # 字体
    plt.rcParams['font.sans-serif'] = ['SimHei']
    # 各块颜色
    colors = ['yellowgreen', 'lightskyblue', 'lightcoral']
    # 饼状图外侧显示的说明文字
    labels = ['未知', '男性', '女性']
    # 离开中心的距离，用于突出显示其中几块
    explode = (0, 0.05, 0.)
    # 标题
    plt.title('微信好友性别比例')
    # 绘制饼图的函数
    plt.pie([unknow, male, female], labels=labels, explode=explode, colors=colors, autopct='%1.1f%%')
    # 输出到文件
    plt.savefig('Sex_Pie.png')
    # 显示饼状图
    plt.show()
```

```
def province_bar(division):
    # 6 个默认的颜色循环主题： deep, muted, pastel, bright, dark, colorblind
    # desat：每种颜色去饱和的比例
    # 设置调色板
    sns.set_palette('deep', desat=.8)
    # 设置 rc 参数显示中文标题，设置字体为 SimHei 显示中文。 rc：run configuration。
    plt.rcParams['font.sans-serif'] = ['SimHei']
    # 从 Province（省份）为山东的列中将 City 对应的城市进行计数。
    # value_counts()是一种查看表格某列中有多少个不同值的快捷方法，并计算每个不同值有在该列中有多少重复值。
    counters = None
    if division == '中国':
        counters = data['Province'].value_counts()
    if division != '中国':
        counters = data[data['Province'] == division]['City'].value_counts()

    counters = counters.sort_values(ascending=False)

    length = 6  # 预设柱状图显示 6 个城市
    if len(counters) > length:
        counters, temp = counters[:length], counters[length:]  # 给变量命名的另外一种方式
        value = sum(temp)  # 将 length 数量之外的城市数量求和
        counters['其他'] = value  # 在 Series(可理解为字典）中插入一组数据（键值对）
    # bar 是柱状图函数，counters.index 表示 X 轴（横坐标）、counters 表示 Y 轴（纵坐标）。
    plt.bar(counters.index, counters)
    # 标题
    plt.title('微信好友区划分布图')
    # 输出到文件
    plt.savefig('City_Bar.png')
    # 显示柱状图
    plt.show()


def signature_word_cloud():
    itchat.auto_login(hotReload=True)
    friends = itchat.get_friends(update=True)[1:]

    # dt = pd.DataFrame(friends)
    # dt.to_csv('ok.csv', encoding='utf_8_sig', header=True, index=True)

    signatures = ''
    for friend in friends:
```

```
            signature = friend['Signature']
            if len(signature) == 0:
                continue
            signature = signature.strip().replace('span', '').replace('class', '').replace('emoji', '')  #
去除无关数据

            signature = re.sub(r'1f(\d.+)', '', signature)
            signatures += ' '.join(jieba.analyse.extract_tags(signature))  # 关键字提取
            signatures += ' '


    # Image.open，读取指定图片。

    im = Image.open('qq.jpg')  # 可替换你喜欢的图片，在当前文件夹下（相对路径）

    # np.array，将读入的 im 转换成背景图数据。
    mask = np.array(im)
    # WordCloud 函数，建立词云对象
    # mask 参数用于设置词云形状，默认的是矩形，可以读入自己选定的图片。margin:
画布偏移，默认 2 像素.
    word_cloud = WordCloud(font_path='simhei.ttf', background_color='white',
max_words=1200, mask=mask, margin=15)
    # generate，向 word_could 这个 WordCloud 对象中加载 signatures（文本内容），对全
部文本进行自动分词（但是对中文支持不好）
    word_cloud.generate(signatures)
    # ImageColorGenerator 函数通过 mask 参数生成词云颜色值
    image_colors = ImageColorGenerator(mask)
    # 用 recolor 方法重置词云颜色为（color_func=image_colors）
    word_clour = word_cloud.recolor(color_func=image_colors)
    # figure 函数中，figsize 表示输出的绘图对象的宽和高、dpi 表示指定绘图对象的分辨
率，即每英寸多少个像素，缺省值为 80。
    plt.figure(figsize=(12, 12), dpi=100)
    # imshow 函数用于对按照样本图片重置颜色的图像进行处理，并显示其格式，但是不
能显示。
    plt.imshow(word_clour)
    # 不显示坐标尺寸
    plt.axis('off')
    # 显示词云图
    plt.show()
    # 输出到文件
    word_cloud.to_file('signatures.png')



while True:
    flag = input('_____\n'
                 'Please press the letter:\n'
```

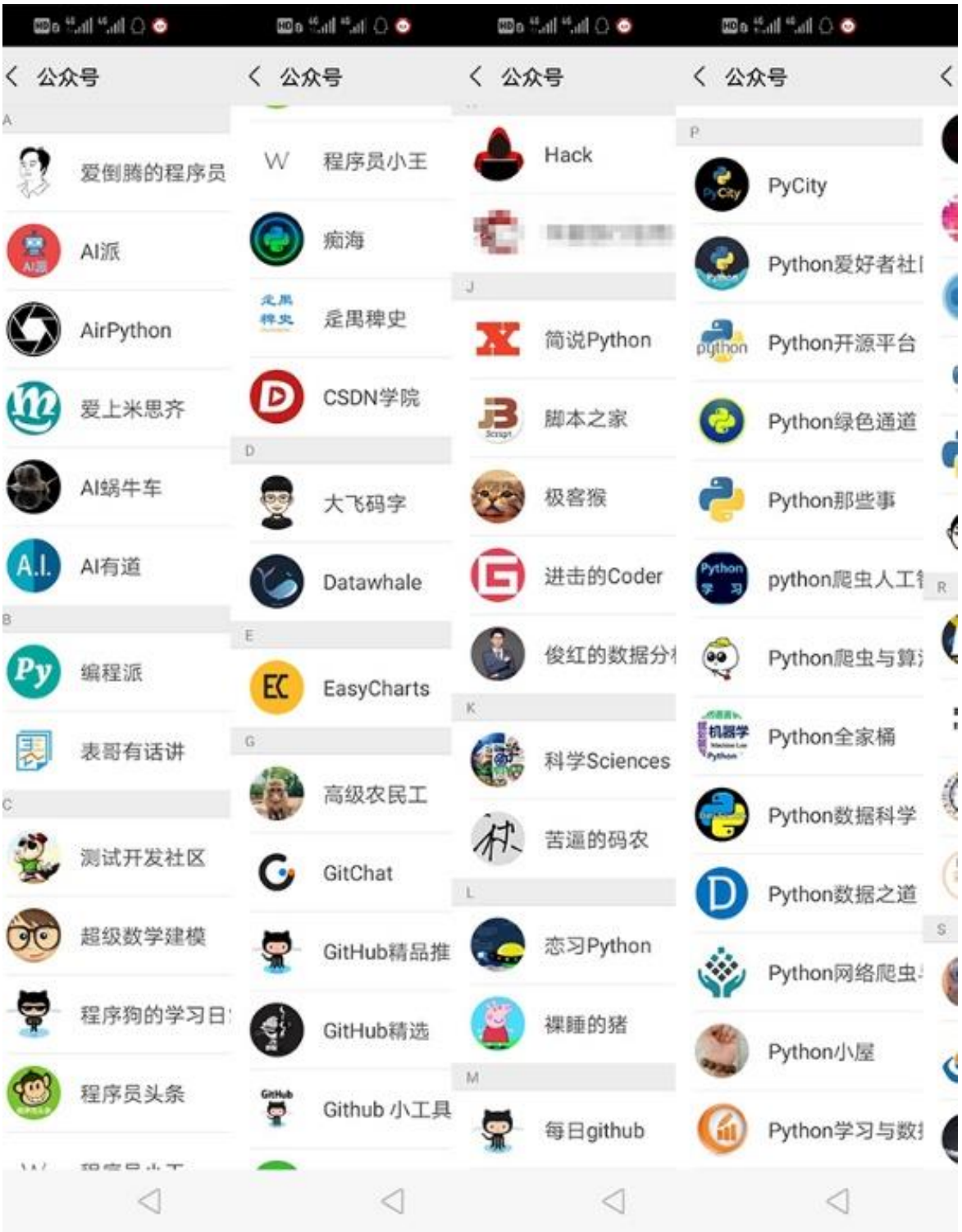
```
                    'Press the \'s\' to sex_pie_chart.\n'
                    'Press the \'p\' to province_bar_chart.\n'
                    'Press the \'w\' to signature_word_cloud.\n'
                    'Press the \'e\' to exit.\n'
                    '_____')
    try:
        if flag == 's':
            sex_pie()

        if flag == 'p':
            division = input('Please input the division:')
            if division == '':
                division = '中国'
            province_bar(division)

        if flag == 'w':
            signature_word_cloud()

        if flag == 'e':
            break
    except ValueError:
        pass

print('\n*****************************************\n'
      '* Thank you for coming, you are the best. *\n'
      '*****************************************')
```

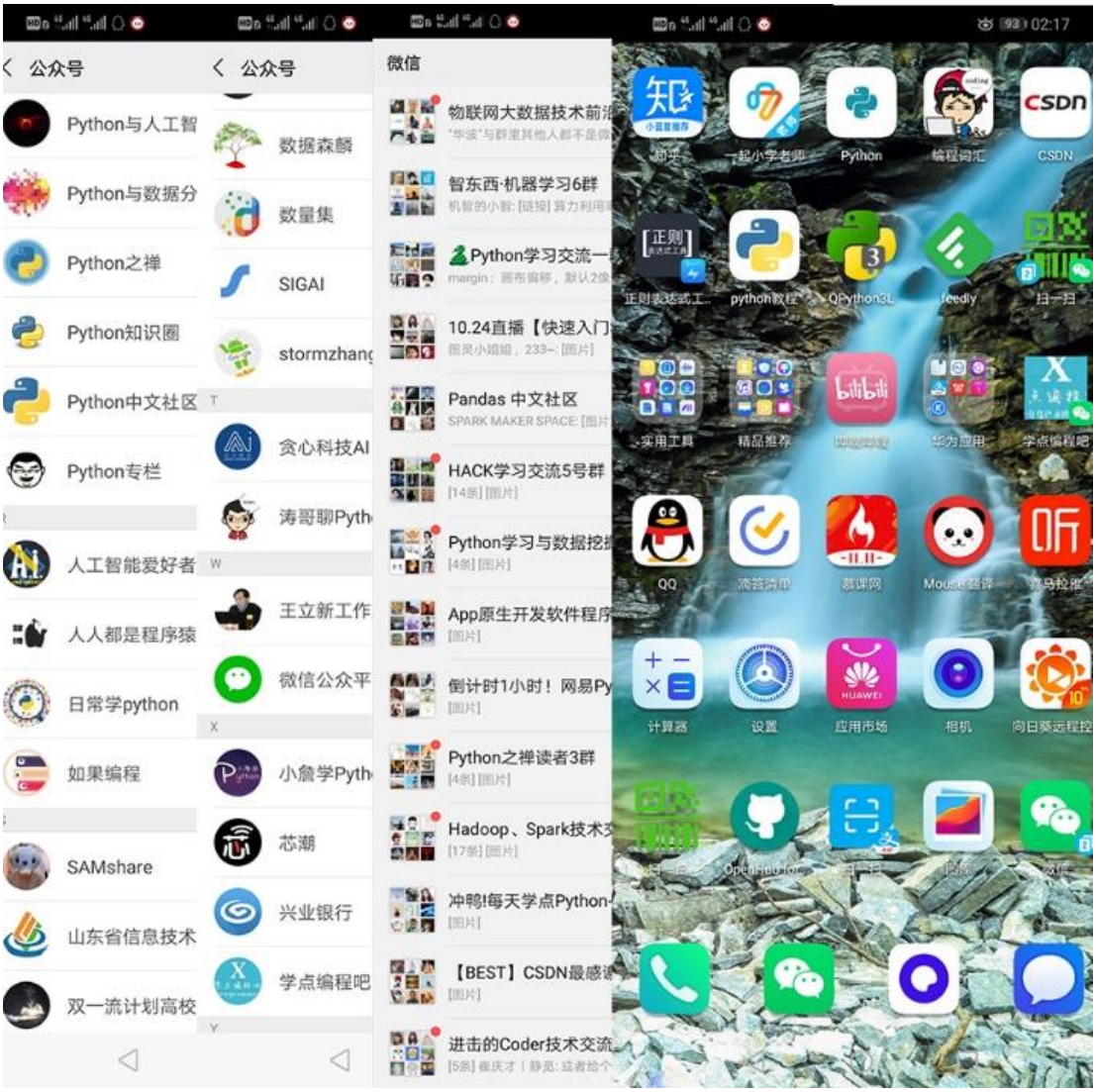
几点个人经验和建议

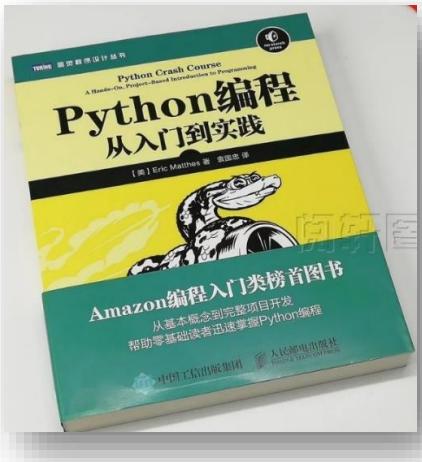
- 我是谁，从哪里来，到哪里去。
- 模块即源文件，内含若干个类。自己写一个 py 源文件，它的名字就是模块名，可以被其他源文件 import。
- 包是一个目录，内含若干个模块和若干个子目录，它的特点是目录下含有__init__.py. 所以一个包也有着一个同名模块。包是一个包含__init__.py 文件的目录，该目录下一定得有这个__init__.py 文件和其它模块或子包。
- 库是相关包的集合。python 中的完成一定功能的代码集合，供用户使用的代码组合。
- 本节课用到的库：itchat, NumPy, Pandas, Matplotlib, Seaborn, wordcloud(要先安装微软 visual c++ 14 或以上全部支持包), jieba, Pillow, openpyxl, xlrd。
- import 包.下级包.下级包.模块名 as 别名
- 本节课目标：一是学习编程是为了理论服务于实践和应用。二是了解各个常见的功能强大的库。三是理解编程的模式，针对需求抽象建模得出算法，代码逻辑实现，调试，完成。四是用逻辑代码实现算法是最重要的。

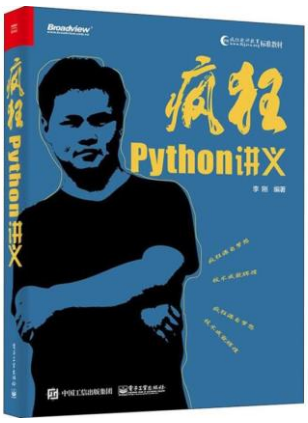
- 注意 PEP8 规范。
- 英文的重要性、官网的重要性、英文官网的重要性，英文官网教程，坚持用英文原版软件不用汉化版，读音不准不要紧。

- 用好搜索引擎。
- github，代码托管，https://github.com/youandpython/WeiHai1029。
- pycharm，详细配置，专业版，配置编码，配置 github，配置 jupyter notebook。
- 代码逻辑和抽象算法不是短时间练出来的，写代码就像数学应用题，时间久了就有感觉了。需要较长时间的学习积累。编程既需要集中思维也需要发散思维。先不要深究，先走一遍。根据问题抽象建模，就是解决问题的框架，取决于对各个库的方法属性的熟悉。python 很宽泛，还需要很多相关周边知识。遇到问题无法解决一定先百度，不要不假思索地问别人，探究的过程后记得牢。带好笔记本和笔，我之前没有机会听课走了很多弯路，大家要把握珍惜。
- 要跟上时代步伐，尤其是微机老师更要创新，最好用 win10。谷歌浏览器自带翻译，绿色。火绒杀毒。射手播放器。向日葵。有道云辞典。其它杂乱的软件不用，不常用的用完接着卸载。
- 要有一个足够快的手机，不限流量的资费套餐，一套实用的便捷的绿色轻量级软件（手机端和电脑端的），有随时有规律的学习方案（碎片时间和非碎片时间，公号，喜马拉雅，B 站，纸质书，本地电子书和视频教程，官方文档），每天写代码读代码，读文章，坚持学英语，尝试针对生活和工作需求写小项目。多加微信群。
- 选对书很重要，就那么点时间，要高效。


公众号
A
爱倒腾的程序员
AI派
AirPython
爱上米思齐
AI蜗牛车
AI有道
B
编程派
表哥有话讲
C
测试开发社区
超级数学建模
程序狗的学习日
程序员头条
公众号
程序员小王
痴海
CSDN学院
D
大飞码字
Datawhale
E
EasyCharts
G
高级农民工
GitChat
GitHub精品推
GitHub精选
Github 小工具
公众号
Hack
J
简说Python
脚本之家
极客猴
进击的Coder
K
科学Sciences
苦逼的码农
L
恋习Python
裸睡的猪
M
每日github
公众号
P
PyCity
Python爱好者社
Python开源平台
Python绿色通道
Python那些事
python爬虫人工
Python爬虫与算
Python全家桶
Python数据科学
Python数据之道
Python网络爬虫
Python小屋
Python学习与数
R
S







关于我：
WeChat: youandpython
QQ: 9674767
E-Mail：zhuchengcomputer@126.com
GitHub: https://github.com/youandpython