

小伙伴计划暑期学习营——零基础Python入门

第四讲：玩转文件

张智帅 电子系

清华大学学生学业与发展指导中心
2019-2020学年夏季学期

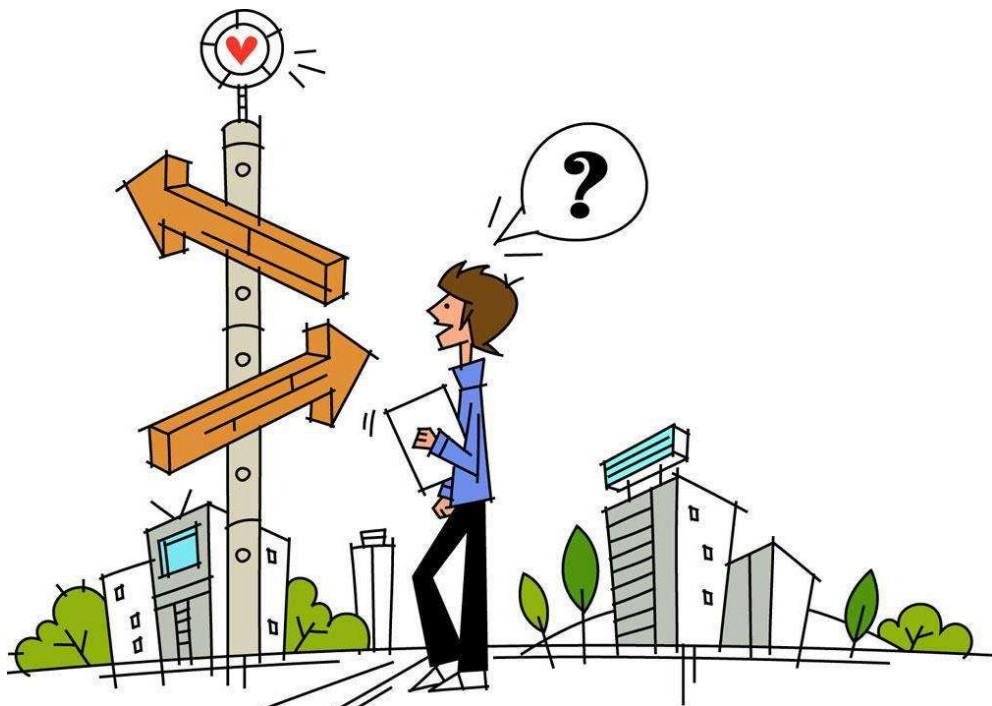
第五讲：玩转文件

■ 路径处理

- 相对路径、绝对路径
- os模块

□ 具体文件处理举例

- 办公文档
 - Word
 - Excel
 - PPT
 - PDF
- 数据文件
 - CSV
 - JSON



工作目录

□ 同一段代码

```
L5_path.py ×  
1  #相对路径  
2  f = open("summersummersummer.txt", "w")  
3  f.write("I wanna go back to Tsinghua.")  
4  f.close()
```

➤ 两种运行地点

```
(base) E:\Testfield\summer_python\Lecture_4>python -u "e:\Testfield\summer_python\Lecture_4\L5_path.py"
```

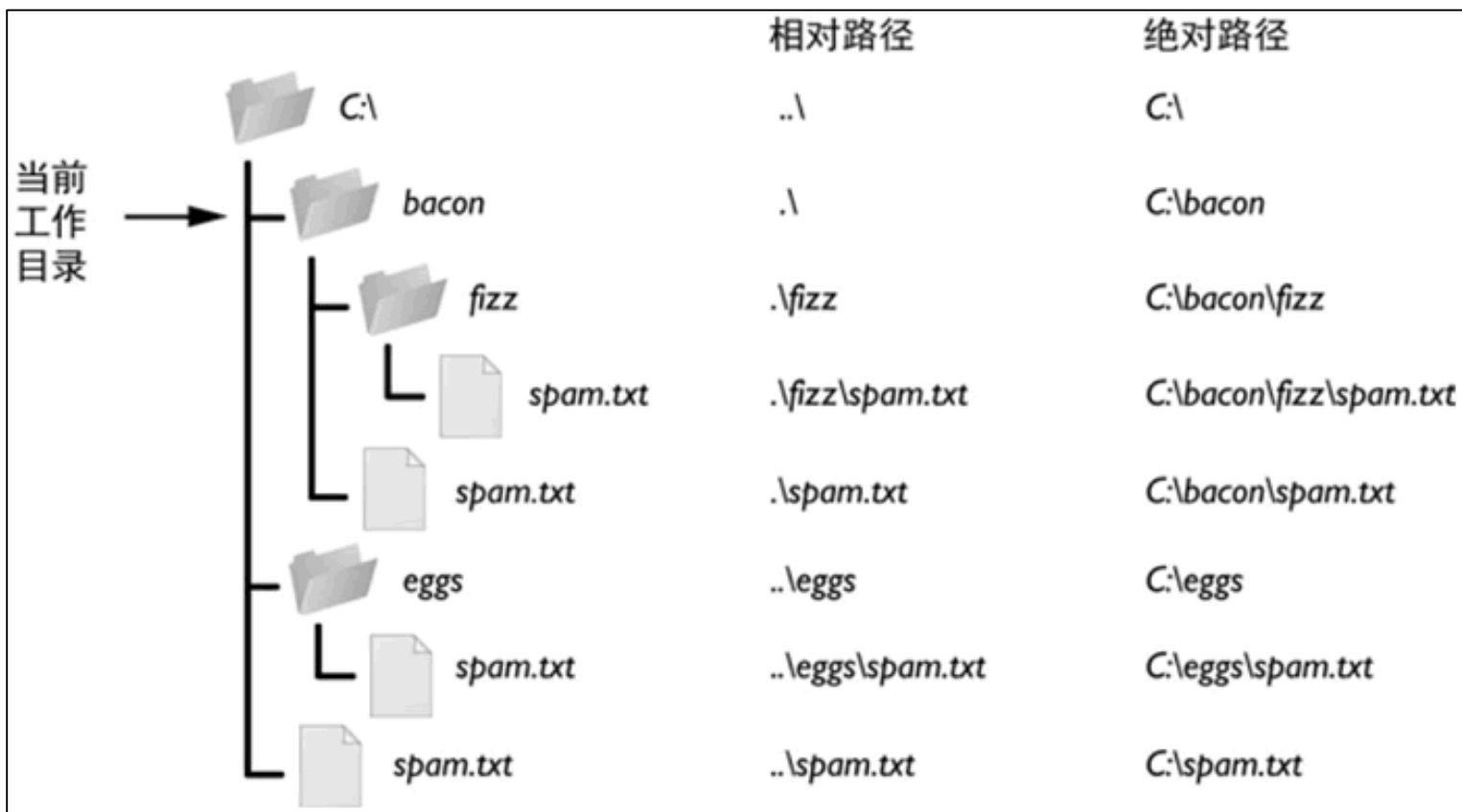
```
(base) C:\Users\hasee>python -u "e:\Testfield\summer_python\Lecture_4\L5_path.py"
```

□ 文件写到哪里去了？

- 当前工作目录 (cwd, current working directory) : 所有没有从根目录开始的文件名或路径, 都假定在当前工作目录下

相对路径、绝对路径

- ❑ 绝对路径：总是从**根目录**开始。Window 系统中以盘符（C：、E：）作为根目录，而 OS X 或者 Linux 系统中以 / 作为根目录。
- ❑ 相对路径：指的是文件相对于**当前工作目录**所在的位置。



斜杠

- ❑ Windows系统：右斜杠 \（反斜杠），或者转义斜杠\\

```
(base) E:\Testfield\summer_python\Lecture_4>
```

- ❑ OS X 或者 Linux 系统：左斜杠 /（除号）

```
arcadia@DESKTOP-JQEVVA5:/mnt/e/Testfield/summer_python/Lecture_4$ pwd  
/mnt/e/Testfield/summer_python/Lecture_4
```

- ❑ Python中不做区分：

```
f1 = open("E:/Testfield/summer_python/Lecture_4/summer.txt", "r")  
f2 = open("E:\\Testfield\\summer_python\\Lecture_4\\summer.txt", "r")  
f3 = open("E:\Testfield\summer_python\Lecture_4\summer.txt", "r")
```

- 跨平台的时候需要考虑兼容性

- 内置模块os提供了操作系统服务，其子模块os.path可以用于路径处理
- 路径处理常用操作：**查看、创建、删除、改名、复制、移动、文件名的拆分与组合.....****根据需要，现用现查**
- 举例：若文件夹不存在则创建

```
import os

folder_name = "txt_samples"
existence = os.path.exists(folder_name) # 判断文件/文件夹是否存在
print(existence)

if not existence: # 若文件夹不存在则创建
    os.mkdir(folder_name)

file_name = os.path.join(folder_name, "summer.txt") # 合并文件名
```

- 官方文档：[os.path — Common pathname manipulations](#)

第五讲：玩转文件

□ 路径处理

■ 具体文件处理举例

➤ 办公文档

- Word
- Excel
- PPT
- PDF

➤ 数据文件

- CSV
- JSON



	社会实践邀请函_阿富汗.docx
	社会实践邀请函_巴勒斯坦.docx
	社会实践邀请函_巴林.docx
	社会实践邀请函_黎巴嫩.docx
	社会实践邀请函_沙特阿拉伯.docx
	社会实践邀请函_伊拉克.docx
	社会实践邀请函_以色列.docx
	社会实践邀请函_约旦.docx
	社会实践邀请函_阿拉伯联合酋长国.docx
	社会实践邀请函_阿曼.docx
	社会实践邀请函_阿塞拜疆.docx
	社会实践邀请函_格鲁吉亚.docx
	社会实践邀请函_卡塔尔.docx
	社会实践邀请函_科威特.docx
	社会实践邀请函_塞浦路斯.docx
	社会实践邀请函_土耳其.docx
	社会实践邀请函_亚美尼亚.docx
	社会实践邀请函_也门.docx

docx文档的格式组成

- docx文件本质上是一个**ZIP文件**
- 主要内容保存为**XML格式**
- document.xml文件包含了文档的主要文本内容

> Testfield > summer_python > Lecture_4 > 背影 - 副本 > word			
名称	修改日期	类型	大小
📁 _rels	2020-7-31 16:24	文件夹	
📁 media	2020-7-31 16:24	文件夹	
📁 theme	2020-7-31 16:24	文件夹	
📄 document.xml		XML 源文件	11 KB
📄 endnotes.xml		XML 源文件	2 KB
📄 fontTable.xml		XML 源文件	2 KB
📄 footer1.xml		XML 源文件	2 KB
📄 footer2.xml		XML 源文件	2 KB
📄 footer3.xml		XML 源文件	2 KB
📄 footnotes.xml		XML 源文件	2 KB
📄 header1.xml		XML 源文件	2 KB
📄 header2.xml		XML 源文件	2 KB
📄 header3.xml		XML 源文件	2 KB
📄 settings.xml		XML 源文件	4 KB
📄 styles.xml		XML 源文件	32 KB
📄 webSettings.xml		XML 源文件	1 KB

docx文档的格式组成

- 段落对象：paragraph
 - 缩进、间距、对齐.....
- 最基本的单位：run
 - 字体、大小、颜色.....

文档 Document

段落 Paragraph

行内元素 Runs

内容 text

字体 font

颜色 color

字号 size

内容 text

```
<p w:rsidR="00251E5F" w:rsidRDefault="00251E5F" w:rsidP="00251E5F">
```

```
<pPr>
```

paragraph

```
<ind w:firstLine="420" />
```

```
</pPr>
```

```
<r>
```

run

```
<rPr>
```

```
<rFonts w:hint="eastAsia" />
```

```
</rPr>
```

text

```
<t>
```

我与父亲不相见已二年余了，我最不能忘记的是他的背影。那年冬天，祖母死了，父亲的差使也交

```
</t>
```

```
</r>
```

```
</p>
```

python-docx

❑ 安装: `pip install python-docx`

❑ 导入: `import docx`

➤ 读取:

```
import docx

file = docx.Document("背影.docx") # 获取文档对象
print("段落数:" + str(len(file.paragraphs))) # 读取文档的段落数

# 输出段落编号及段落内容的前三十个字
for i, paragraph in enumerate(file.paragraphs):
    print("第" + str(i) + "段: " + paragraph.text[:30])
```

段落数:12

第0段: 背影

第1段: 我与父亲不相见已二年余了,我最不能忘记的是他的背影。那年冬天

第2段: 回家变卖典质,父亲还了亏空;又借钱办了丧事。这些日子,家中光

第3段: 到南京时,有朋友约去游逛,勾留了一日;第二日上午便须渡江到浦

第4段: 我们过了江,进了车站。我买票,他忙着照看行李。行李太多了,得

第5段: 我说道,“爸爸,你走吧。”他望车外看了看,说,“我买几个橘子

python-docx

□ 写入

```
import docx
from L4_info import * # 从L4_info文件导入写好的文本

# 创建文档对象
doc = docx.Document()

# 添加标题
doc.add_heading(title, level=1)

# 添加段落
doc.add_paragraph(receiver)
doc.add_paragraph(text)
doc.add_paragraph(sender)
doc.add_paragraph(date)

# 保存文档（保存时请确认该文档是关闭的，否则会出权限错误）
doc.save("Output/社会实践邀请函_仿制.docx")
```

- 官方教程: <https://python-docx.readthedocs.io/en/latest/user/quickstart.html>
- 官方文档: <https://python-docx.readthedocs.io/en/latest/api/document.html>

□ 调整paragraph和run的样式.....**根据需要，现用现查**

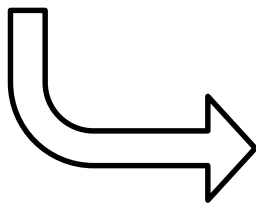
关于邀请华清大学师生赴利比亚进行暑期社会实践的函

共青团华清大学委员会：

获悉华清大学正在组织优秀学生利用暑假赴海外国家进行社会实践的活动，作为在利比亚的华人餐馆，我们诚挚的邀请以华清大学画画学院张小明教授带队的师生代表团共计 3 人来利比亚进行社会实践，画画。

利比亚华人餐馆有限公司

2020 年 07 月 31 日



关于邀请华清大学师生赴利比亚进行暑期社会实践的函

共青团华清大学委员会：

获悉华清大学正在组织优秀学生利用暑假赴海外国家进行社会实践的活动，作为在利比亚的华人餐馆，我们诚挚的邀请以华清大学画画学院张小明教授带队的师生代表团共计 3 人来利比亚进行社会实践，画画。

利比亚华人餐馆有限公司

2020 年 07 月 31 日

第五讲：玩转文件

□ 路径处理

■ 具体文件处理举例

➤ 办公文档

- Word
- Excel
- PPT
- PDF

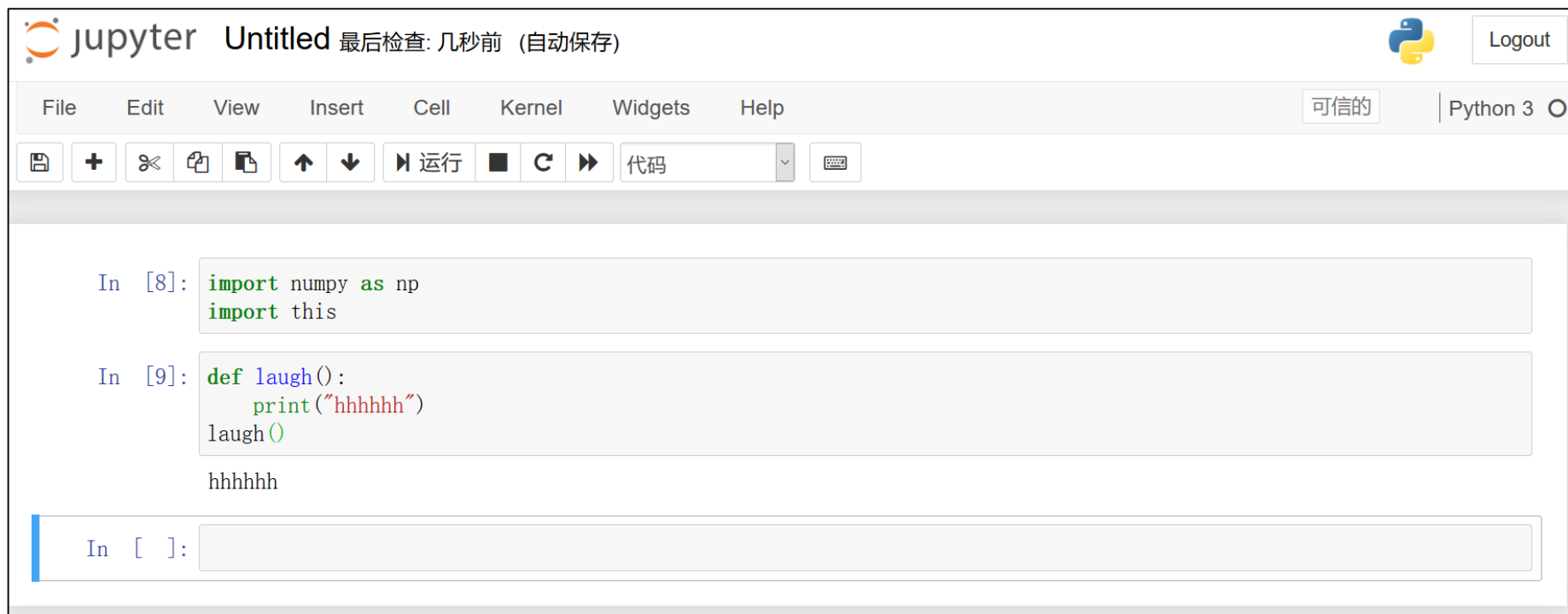
➤ 数据文件

- CSV
- JSON



Jupyter notebook

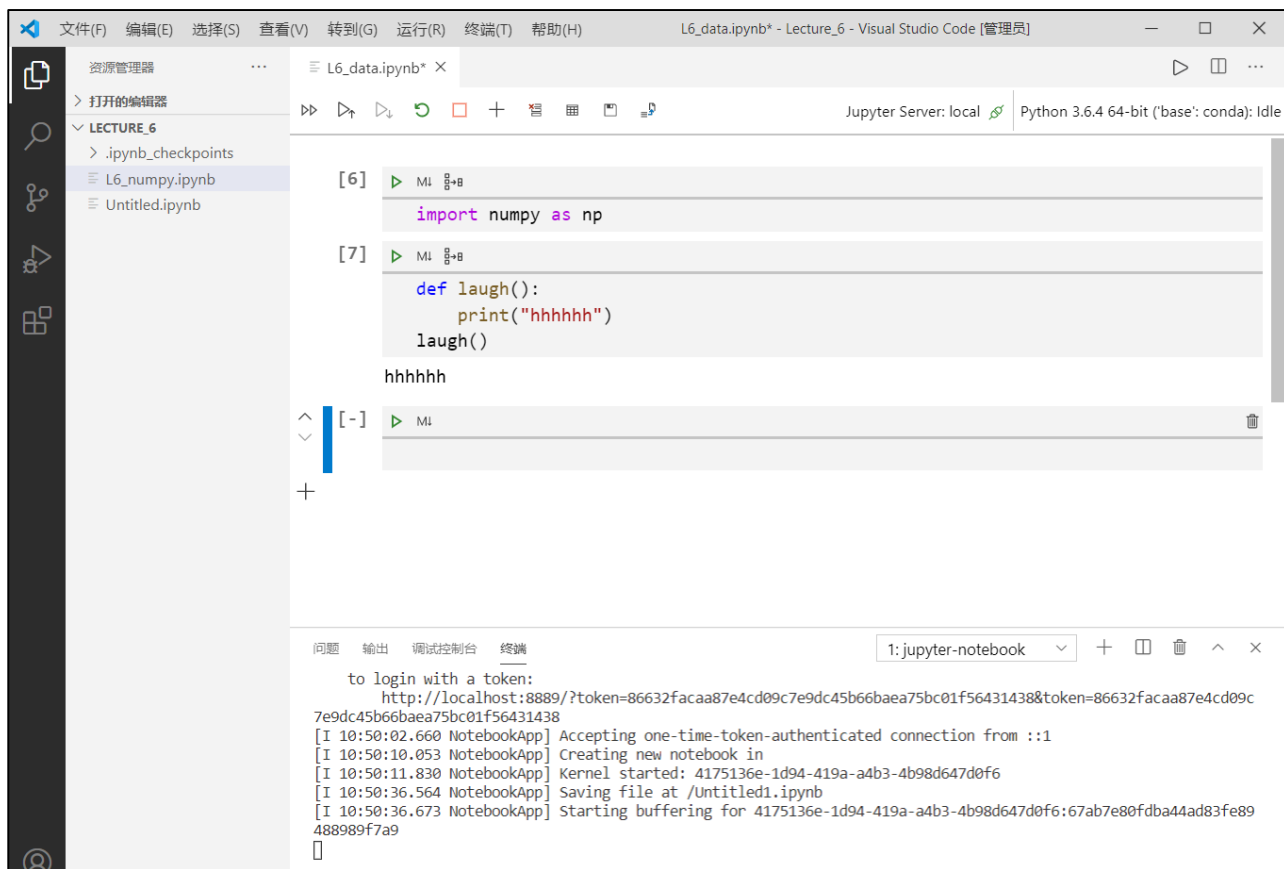
- 基于网页，交互计算。
- 可以运行代码，并且**展示结果**❤❤❤



- [Jupyter Notebook介绍、安装及使用教程](#)
- [Jupyter快捷键总结](#)

Jupyter notebook

- ❑ Vscode支持原生notebook❤️❤️❤️
- ❑ 把文件保存为.ipynb格式即可自动识别



Excel

□ 常见第三方库

[illegible]

xlwings, 让excel飞起来

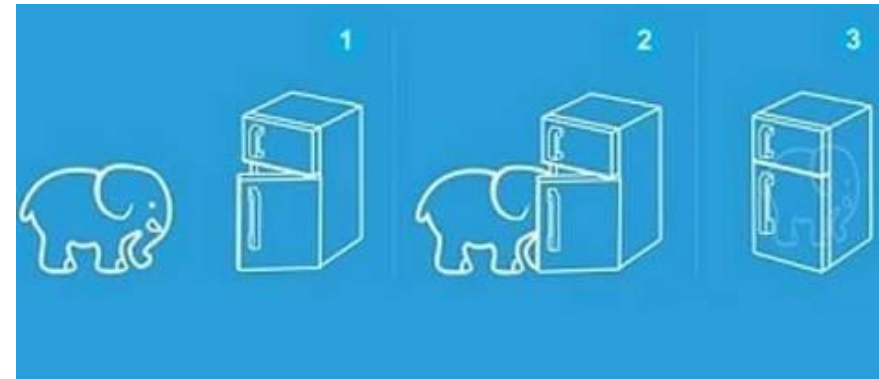
- ❑ 安装: `pip install xlwings`
- ❑ 导入: `import xlwings as xw`
- ❑ 使用步骤: **打开应用-操作-退出应用**

```
[52] ▶ MI
# 若无法导入, 则pip install xlwings
import xlwings as xw

[56] ▶ MI
"""
新建一个表格应用
visible: 是否可见
add_book: 是否新建工作簿
"""
app = xw.App(visible=True, add_book=False)

[57] ▶ MI
# 打开工作表
wb = app.books.open('社会实践报名表.xlsx')

# 打开工作簿
sht = wb.sheets["sheet1"]
```



```
▶ MI
# 使用结束一定要记得关闭
wb.save() #保存工作表
wb.close() #关闭工作簿
app.quit() #退出表格应用
```

xlwings, 让excel飞起来

□ 读取：按照单元格坐标

[73] ▶ MI

```
# 读取特定单元格
print(sht.range("A1"))
print(sht.range("A1").value) #读取值
print(sht["A1"].value)
print(sht["A1:B2"].value)
print(sht["$A$1:$B$2"].value)
```

<Range [社会实践报名表.xlsx]Sheet1!\$A\$1>

目的地

目的地

[['目的地', '描述'], ['阿富汗', '第6公司']]

[['目的地', '描述'], ['阿富汗', '第6公司']]

□ 也可以直接获得已使用的范围

[83] ▶ MI


```
sht.used_range.shape
```


(19, 5)


xlwings + python-docx


□ 读取Excel表格，批量生成Word文档


```
15 > def read_info_xlwings(filename): ...
32
33
34 > def write_invitation(target, description, department, number, plan, filename): ...
89
90
91 > def mkdir(folder_name): ...
96
97
98 # 读取 excel
99 file_source = "Data_samples/社会实践报名表.xlsx"
100 countries, descriptions, departments, numbers, plans = read_info_xlwings(file_source)
```


 社会实践邀请函_阿富汗.docx


 社会实践邀请函_巴勒斯坦.docx


 社会实践邀请函_科威特.docx


 社会实践邀请函_土耳其.docx


 社会实践邀请函_以色列.docx


 社会实践邀请函_阿拉伯联合酋长国.docx


 社会实践邀请函_巴林.docx


 社会实践邀请函_黎巴嫩.docx


 社会实践邀请函_亚美尼亚.docx


 社会实践邀请函_约旦.docx


 社会实践邀请函_阿曼.docx


 社会实践邀请函_格鲁吉亚.docx


 社会实践邀请函_塞浦路斯.docx

 社会实践邀请函_也门.docx

 社会实践邀请函_阿塞拜疆.docx

 社会实践邀请函_卡塔尔.docx

 社会实践邀请函_沙特阿拉伯.docx

 社会实践邀请函_伊拉克.docx

xlwings, 让excel飞起来

□ 修改：读取+赋值

[84] ▶ M1

直接赋值

sht["A2"].value = "阿富汗" #一一对应

□ 范围赋值

```
print(countries)
```

```
['阿富汗', '伊拉克', '约旦', '黎巴嫩', '以色列', '巴勒斯坦', '沙特阿拉伯', '巴林',  
'卡塔尔', '科威特', '阿拉伯联合酋长国', '阿曼', '也门', '格鲁吉亚', '亚美尼亚',  
'阿塞拜疆', '土耳其', '塞浦路斯']
```

[87] ▶ M1

默认按照行赋值

sht["A2"].value = countries

[88] ▶ M1

范围赋值

sht["B2:Z2"].value = "" # 赋值为空字符串，即把上述行赋值清空

[89] ▶ M1

按列赋值

sht["A2"].options(transpose=True).value = countries #加一句转置

➤ 官方教程: <https://docs.xlwings.org/en/stable/quickstart.html>

➤ 官方文档: <https://docs.xlwings.org/en/stable/api.html>

pandas

- ❑ 安装: `pip install pandas`
- ❑ 导入: `import pandas as pd`
- ❑ 按照列名读取

```
[5] ▶ MI  
for i in range(10):  
    print(df['目的地'].values[i])
```

阿富汗
伊拉克
约旦
黎巴嫩
以色列
巴勒斯坦
沙特阿拉伯
巴林
卡塔尔
科威特

```
[2] ▶ MI  
# 若无法导入, 则pip install pandas  
import pandas as pd  
  
[4] ▶ MI  
df = pd.read_excel("社会实践报名表.xlsx")  
df
```

	目的地	描述	院系名	人数	计划
0	阿富汗	第5公司	航院	20	篮球
1	伊拉克	第2公司	自动化	3	唱
2	约旦	第7公司	电子	3	唱
3	黎巴嫩	第8公司	航院	12	跳
4	以色列	第8公司	工物	10	rap
5	巴勒斯坦	第2公司	航院	17	唱
6	沙特阿拉伯	第2公司	土木	7	唱
7	巴林	第7公司	电子	9	唱
8	卡塔尔	第8公司	土木	15	篮球
9	科威特	第5公司	航院	19	跳
10	阿拉伯联合酋长国	第7公司	航院	8	跳
11	阿曼	第7公司	工物	20	唱

- ❑ 函数`read_excel`的官方文档: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_excel.html

pandas

- ❑ 不能在原 excel 中修改，只能读取、写入新表格

```
[7] ▶ M4  
# 写入 excel  
df.to_excel("社会实践报名表_pandas.xlsx")
```

- ❑ 读写速度快
- 不需要打开Excel应用程序

```
[Running] python -u "e:\Testfield\summer_python\Lecture_4\L4_batch_invitation.py"  
[Done] exited with code=0 in 9.967 seconds read_info_xlwings  
[Running] python -u "e:\Testfield\summer_python\Lecture_4\L4_batch_invitation.py"  
[Done] exited with code=0 in 3.335 seconds read_info_pandas
```

- ❑ 更多pandas内容，见第五讲

第五讲：玩转文件

□ 路径处理

■ 具体文件处理举例

➤ 办公文档

- Word
- Excel
- PPT
- PDF

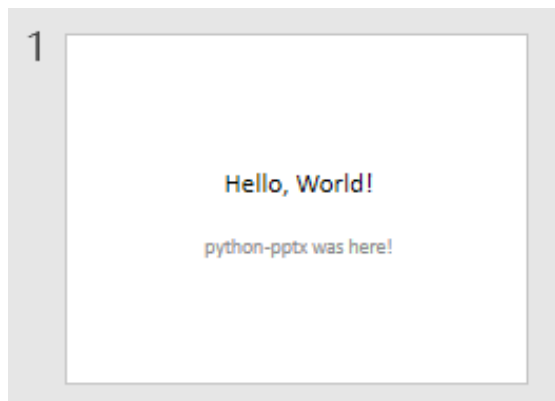
➤ 数据文件

- CSV
- JSON



python-pptx*

- ❑ 安装: `pip install python-pptx`
- ❑ 导入: `import pptx`



```
import pptx

pre = pptx.Presentation() # 创建 PPT 对象

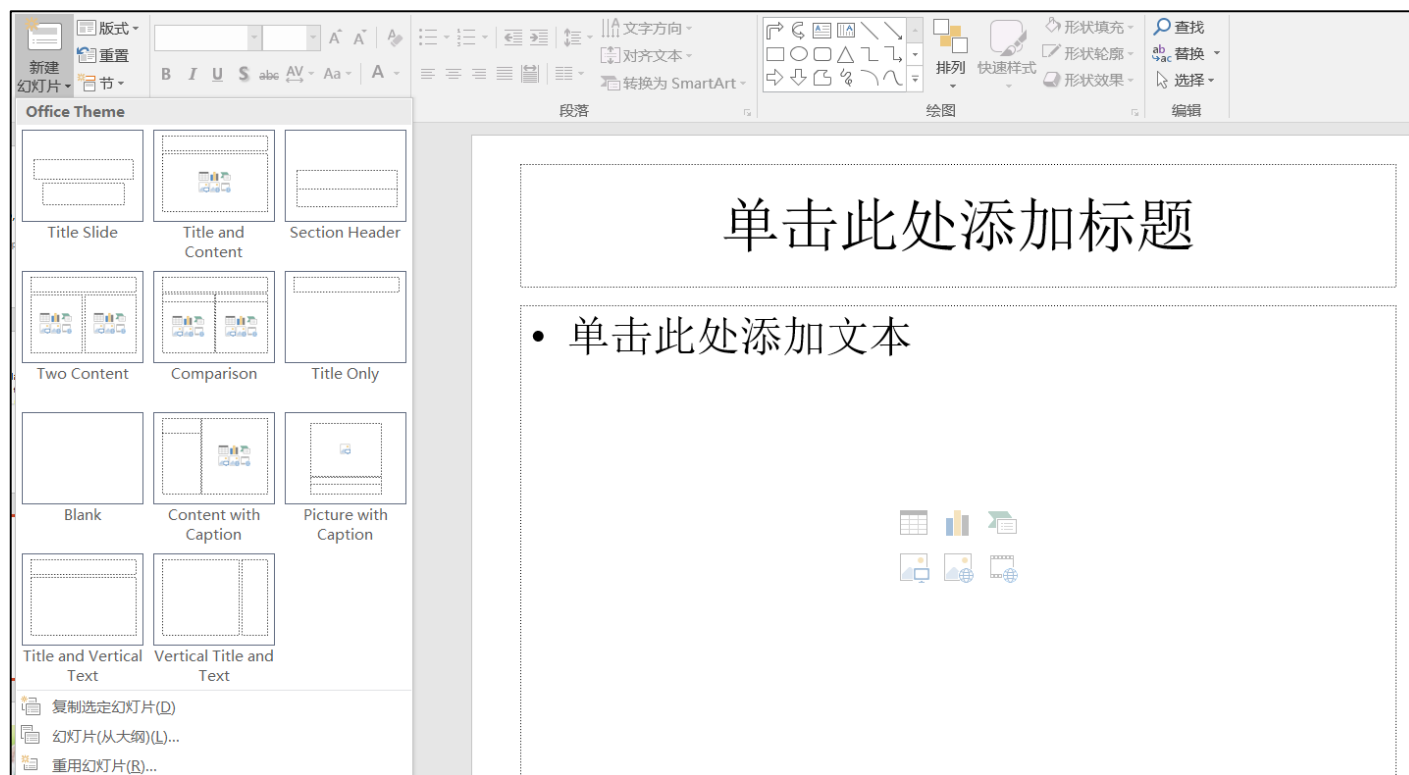
# 第一页
title_slide_layout = pre.slide_layouts[0] # 主题样式序号
slide = pre.slides.add_slide(title_slide_layout) # 增加到 PPT 对象中

title = slide.shapes.title
subtitle = slide.placeholders[1]
title.text = "Hello, World!"
subtitle.text = "python-pptx was here!"
```

- 官方教程: <https://python-pptx.readthedocs.io/en/latest/user/quickstart.html>
- 官方文档: <https://python-pptx.readthedocs.io/en/latest/#api-documentation>
- 超详细中文教程: [python-pptx 实践 1: 创建 PPT 文档](#)

python-pptx*

- 根据主题样式创建页面: `pre.slide_layouts[0]~[10]`
- 添加到PPT对象中: `pre.slides.add_slide(title_slide_layout)`
- 通过占位符添加内容: `slide.placeholders[0]`



第五讲：玩转文件

□ 路径处理

■ 具体文件处理举例

➤ 办公文档

- Word
- Excel
- PPT
- PDF

➤ 数据文件

- CSV
- JSON



福昕PDF编辑器 个人版

永久免费试用的高效PDF编辑器，一键搞定PDF编辑、合并、转换、水印

立即下载

开通会员

用户账号：

购买服务：

编辑特权包

福昕会员

可选套餐：

推荐

¥ 198 | 12月

¥ 396 | 2年

¥ 594 | 3年

服务说明：

一个账号支持3台电脑浮动使用

PyPDF2

▣ 三剑客：PdfFileMerger; PdfFileReader, PdfFileWriter

```
# pip install PyPDF2
import PyPDF2

# 创建一个 PdfFileReader 对象
merger = PyPDF2.PdfFileMerger()

# 打开文件
input1 = open("Data_samples\程序员健康指南.pdf", "rb")

# 获取范围内的页数，增加到 merger 中
merger.append(fileobj=input1, pages=(72, 84)) # 73页-84页
merger.append(fileobj=input1, pages=(104, 119)) # 105页-119页

# 输出到文件中
output = open("Output/Chapter_5_7.pdf", "wb")
merger.write(output)
```

- 官方文档: <https://pythonhosted.org/PyPDF2/index.html>
- 超详细中文教程: [Python应用【PDF处理-pypdf2】](#)

第五讲：玩转文件

□ 路径处理

■ 具体文件处理举例

➤ 办公文档

➤ Word

➤ Excel

➤ PPT

➤ PDF

➤ 数据文件

➤ CSV

➤ JSON



JSON

- ❑ JSON (JavaScript Object Notation, JS 对象简谱) 是一种轻量级的纯文本格式，用于存储和表示数据
 - 完全独立于编程语言
 - 易于人类阅读和机器解析
 - 网络传输效率高
- ❑ 内置模块json，读&写

```
[2] ▶ MI
import json

[3] ▶ MI
# 从 json 文件 "JSON_sample.json" 读取数据，存入变量 data
with open("JSON_sample.json", "rb") as f:
    data = json.load(f)

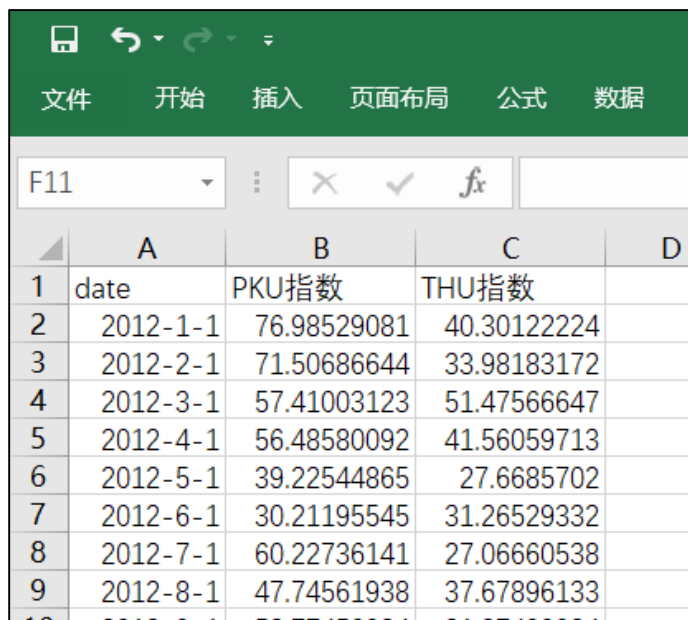
[-] ▶ MI
# 把字典对象 data 写入 json 文件 "names.json"
with open("names.json", "w") as f:
    json.dump(data, f)
```

```
es.json  JSON_sample.json X  Untitled.ipynb  L4_pypdf2.py
1  {
2      "0": {
3          "msg": "",
4          "data": {
5              "count": 241,
6              "previous": null,
7              "results": [
8                  {
9                      "to_user": 14442848,
10                     "liked": 0,
11                     "user_id": 31084077,
12                     "is_essence": 0,
13                     "replies": [],
14                     "deleted": false,
15                     "user_info": {
16                         "user_id": 31084077,
17                         "name": "",
18                         "school_number": "",
19                         "role": 5,
20                         "avatar": "http://pbp38mcp7.",
21                         "nickname": ""
22                     },
```

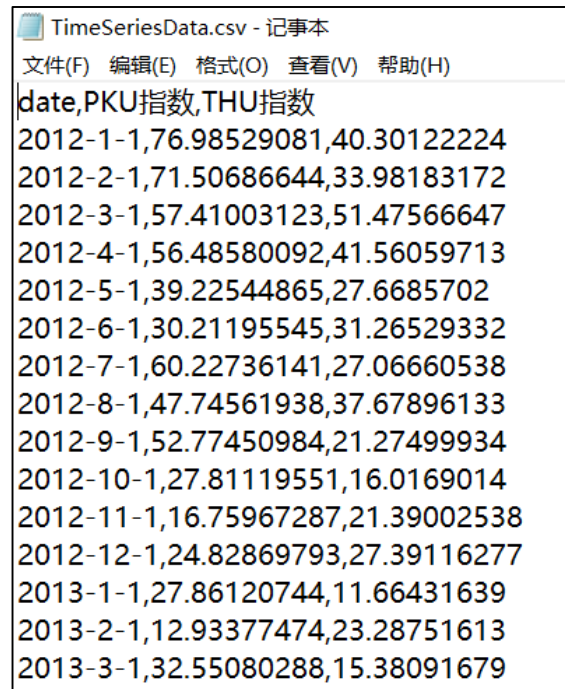
CSV

❑ 逗号分隔值文件（Comma-Separated Values, CSV）

- 以纯文本形式存储**表格数据**（数字和文本）
- CSV文件由任意数目的记录组成，每条记录之间以某种换行符分隔
- 每条记录由字段组成，字段间的分隔符是其它字符或字符串，最常见的是逗号或制表符



	A	B	C	D
1	date	PKU指数	THU指数	
2	2012-1-1	76.98529081	40.30122224	
3	2012-2-1	71.50686644	33.98183172	
4	2012-3-1	57.41003123	51.47566647	
5	2012-4-1	56.48580092	41.56059713	
6	2012-5-1	39.22544865	27.6685702	
7	2012-6-1	30.21195545	31.26529332	
8	2012-7-1	60.22736141	27.06660538	
9	2012-8-1	47.74561938	37.67896133	



```
TimeSeriesData.csv - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
date,PKU指数,THU指数
2012-1-1,76.98529081,40.30122224
2012-2-1,71.50686644,33.98183172
2012-3-1,57.41003123,51.47566647
2012-4-1,56.48580092,41.56059713
2012-5-1,39.22544865,27.6685702
2012-6-1,30.21195545,31.26529332
2012-7-1,60.22736141,27.06660538
2012-8-1,47.74561938,37.67896133
2012-9-1,52.77450984,21.27499934
2012-10-1,27.81119551,16.0169014
2012-11-1,16.75967287,21.39002538
2012-12-1,24.82869793,27.39116277
2013-1-1,27.86120744,11.66431639
2013-2-1,12.93377474,23.28751613
2013-3-1,32.55080288,15.38091679
```

- ❑ 函数read_csv的官方文档: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html

第四讲总结

路径处理

绝对路径与相对路径

os模块 ⊖ eg. 若不存在文件夹则创建

具体文件处理举例

办公文档 ⊖

Word ⊖

python-docx

Excel ⊖

xlwings

pandas

PPT ⊖

python-pptx

PDF ⊖

PyPDF2

数据文件 ⊖

CSV ⊖

pandas

JSON ⊖

json

eg. 批量仿造邀请函

课后练习

1. **阅读并跟做**python-docx官方教程: <https://python-docx.readthedocs.io/en/latest/user/quickstart.html>
2. **阅读并跟做**xlwings官方教程:
<https://docs.xlwings.org/en/stable/quickstart.html>

反馈问卷

