requests模块入门

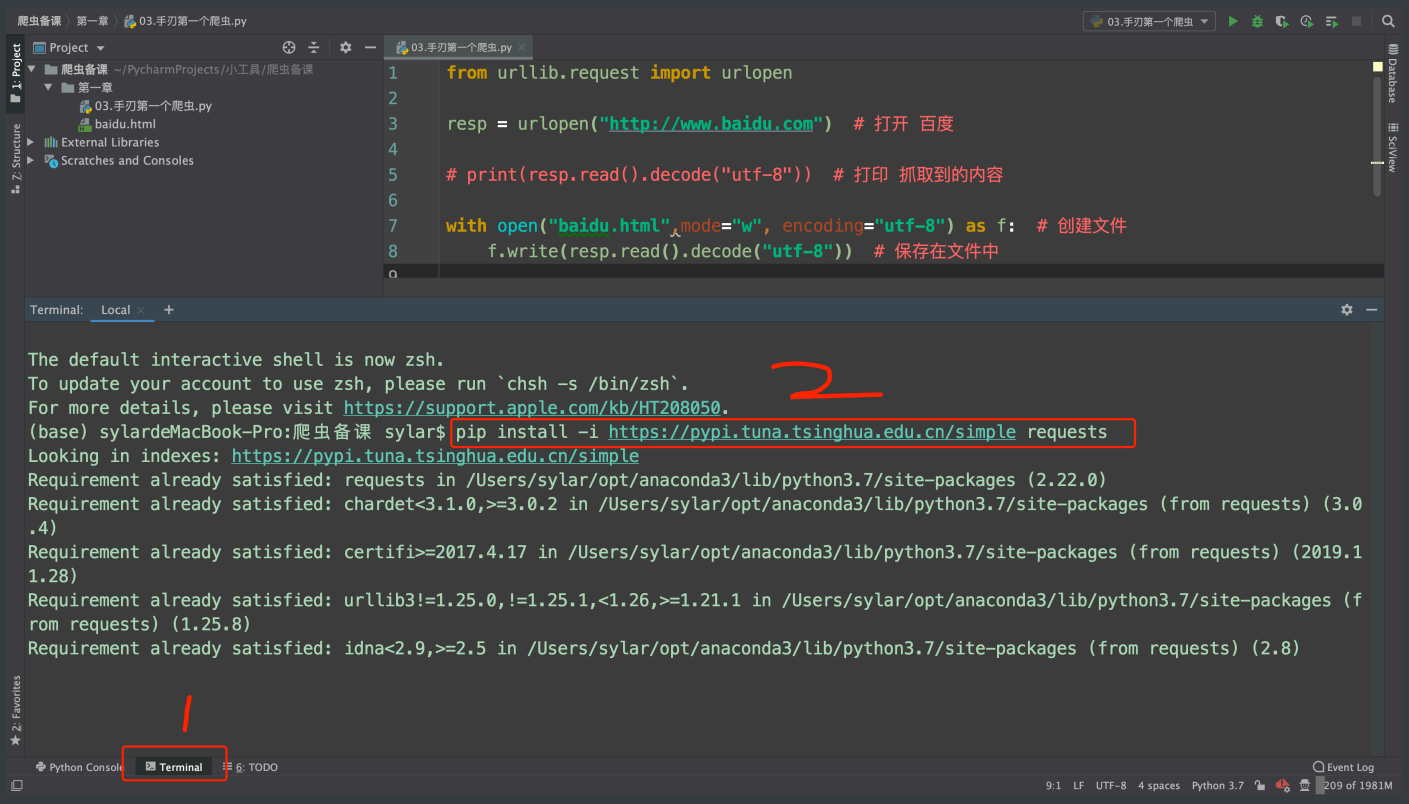
在前面小节中, 我们使用urllib来抓取页面源代码. 这个是python内置的一个模块. 但是, 它并不是我们常用的爬虫工具. 常用的抓取页面的模块通常使用一个第三方模块requests. 这个模块的优势就是比urllib还要简单, 并且处理各种请求都比较方便.

既然是第三方模块, 那就需要我们对该模块进行安装, 安装方法:

```
1 pip install requests
```

如果安装速度慢的话可以改用国内的源进行下载安装.

```
1 pip install -i
  https://pypi.tuna.tsinghua.edu.cn/simple requests
```



OK. 接下来我们来看看requests能带给我们什么?

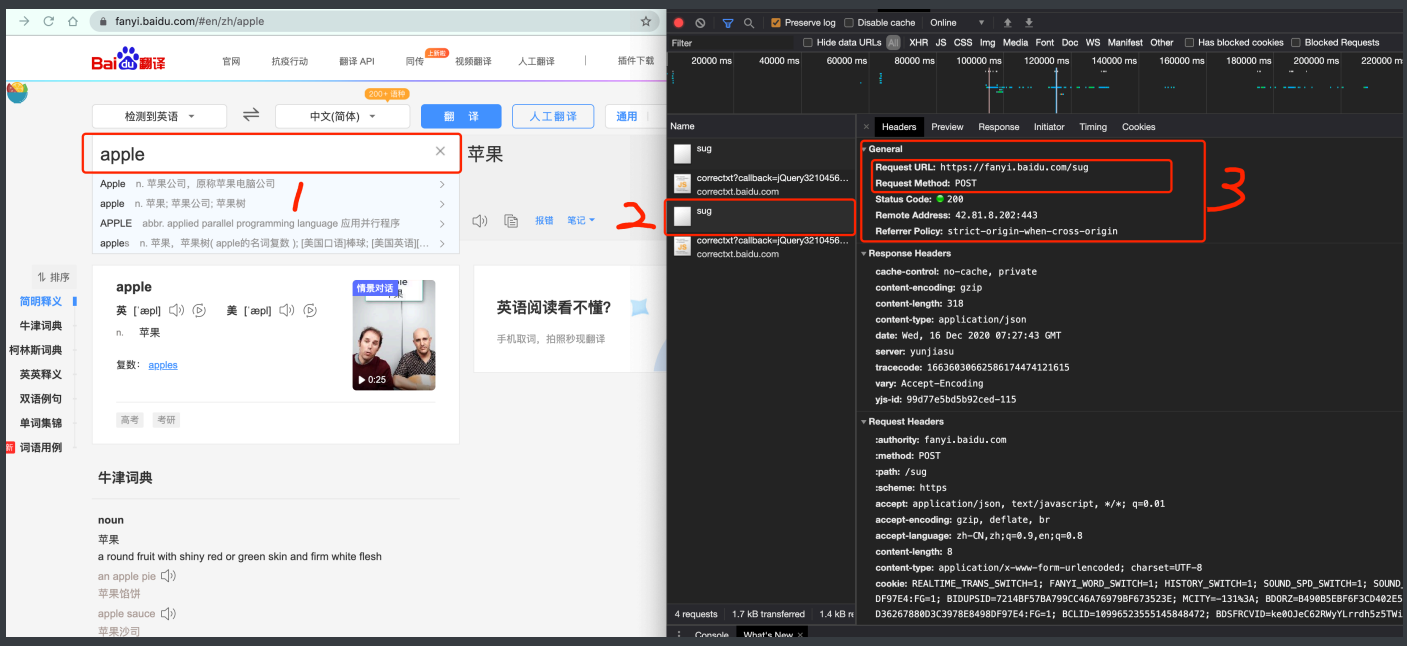
先拿sogou开刀试试.

```python
# 案例1. 抓取搜狗搜索内容
kw = input("请输入你要搜索的内容:")
response = requests.get(f"https://www.sogou.com/web?query={kw}")  # 发送get请求
# print(response.text)  # 直接拿结果(文本)

with open("sogou.html", mode="w", encoding="utf-8") as f:
    f.write(response.text)
```

接下来, 我们看一个稍微复杂那么一丢丢的, 百度翻译~

注意百度翻译这个url不好弄出来. 记住, 在输入的时候, 关掉各种输入法, 要用英文输入法, 然后不要回车. 就能看到这个sug了

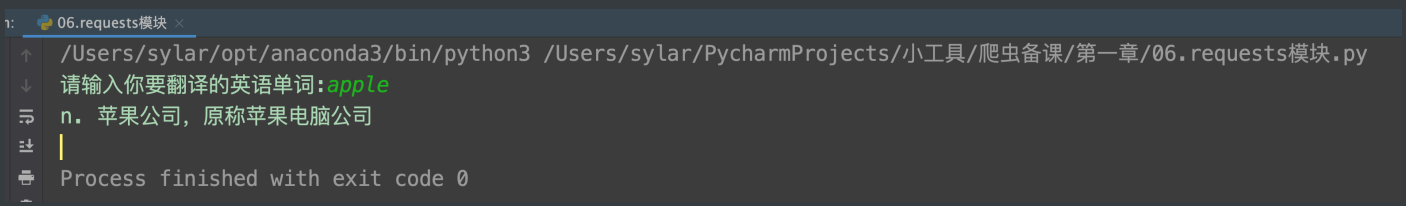


```python
# 案例2.抓取百度翻译数据

# 准备参数
kw = input("请输入你要翻译的英语单词:")
dic = {
    "kw": kw  # 这里要和抓包工具里的参数一致.
}
# 请注意百度翻译的sug这个url. 它是通过post方式进行提交的. 所以我们也要模拟post请求
resp = requests.post("https://fanyi.baidu.com/sug", data=dic)

# 返回值是json 那就可以直接解析成json
resp_json = resp.json()
# {'errno': 0, 'data': [{'k': 'Apple', 'v': 'n. 苹果公司, 原称苹果电脑公司'....
print(resp_json['data'][0]['v'])  # 拿到返回字典中的内容

```



是不是很顺手呢? 还有一些网站在进行请求的时候会校验你的客户端设备型号. 比如, 我们抓取豆瓣电影

```python
# 案例3: 抓取豆瓣电影
url = 'https://movie.douban.com/j/chart/top_list'
param = {
    'type': '24',
    'interval_id': '100:90',
    'action':'',
    'start': '0',#从库中的第几部电影去取
    'limit': '20',#一次取出的个数
}
headers = {
    'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/72.0.3626.121 Safari/537.36'

}
response = requests.get(url=url,params=param,headers=headers)

list_data = response.json()

fp = open('./douban.json','w',encoding='utf-8')
json.dump(list_data,fp=fp,ensure_ascii=False)
print('over!!!')
```

OK~ 本章和本小节的内容就这么多了. 简单回顾一下本章内容

1. 爬虫就是写程序去模拟浏览器用来抓取互联网上的内容
2. python中自带了一个urllib提供给我们进行简易爬虫的编写
3. requests模块的简单使用, 包括get, post两种方式的请求. 以及User-Agent的介绍.