数据获取

数据来源：全国一体化在线政务服务平台国家文物局综合行政管理平台，共5294条数据

发现页面中的表格数据实际上请求了一个后端api接口，通过指定`pageNumber`、`pageSize`获取对应的数据项，为了方便处理，写了一个爬虫来获取数据（按照一页80条数据，共67页进行请求，请求间隔2s）：

```
import requests
import time
import json

headers = {
    "Accept": "*/*",
    "Accept-Encoding": "gzip, deflate, br",
    "Accept-Language": "zh-CN,zh;q=0.9",
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) \
    Chrome/78.0.3904.108 Safari/537.36",
    'Content-Type': 'application/json'
}

result = []

def getdata(page):
    payload =
{"pageNumber":page,"pageSize":80,"keyWord":None,"batchNum":None,"ncrpuScTypeCode
":None,"publishWayCode":None}
    r = requests.post('http://gl.ncha.gov.cn:9200/api/portal/dataNcrpuSc/find',
data=json.dumps(payload), headers=headers)
    return r.json()

def process(response):
    for data in response['data']:
        thing = {}
        thing['id'] = data['serialNum']
        thing['name'] = data['name']
        thing['batch'] = data['batchNumName']
        thing['type'] = data['ncrpuScTypeName']
        thing['address'] = data['address']
        thing['publish'] = data['publishWayName']
        thing['age'] = data['age']
        result.append(thing)

for i in range(1,68):
    process(getdata(i))
    print(i)
    time.sleep(2)

with open('result.json', 'w', encoding='utf-8') as f:
    json.dump(result, f, ensure_ascii=False, indent=4)
```

接口返回的原始数据包含如下字段：


```
{
    "gmtModified" : "0",
    "serialNum" : "1",
    "address" : "广东省广州市三元里",
    "batchNumName" : "第一批",
    "publishWayCode" : "1",
    "ncrpuScTypeName" : "革命遗址及革命纪念建筑物",
    "batchNum" : "1",
    "remark" : null,
    "classNum" : "1",
    "gmtCreate" : "2022-07-08 12:01:15.928386+08",
    "ncrpuScTypeCode" : "105",
    "name" : "三元里平英团遗址",
    "id" : "i7wGwRpBTqJzl5Wxqop",
    "publishWayName" : "文物保护单位",
    "age" : "1841年"
}
```


这里只选取文物保护单位名称、批次、地址、文物类型、公布类型以及文物所处年代进行分析

数据预处理

这部分主要完成两件工作：

1. 依据批次和爬取下来的次序自己制定一个id，便于后续处理数据
2. 补充地理坐标信息

在补充地理坐标信息时，使用到了高德开放平台的地理编码API，个人用户每日5000次免费调用额度，刚好两天时间可以处理完所有数据，代码如下：

```
import requests
import json
import time

headers = {
    "Accept": "*/*",
    "Accept-Encoding": "gzip, deflate, br",
    "Accept-Language": "zh-CN,zh;q=0.9",
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) \
    Chrome/78.0.3904.108 Safari/537.36",
    'Content-Type': 'application/json'
}

def pre_process():
    with open('result.json', 'r', encoding='utf-8') as f:
        data = json.load(f)
    reflect = {
        '一': '1',
        '二': '2',
        '三': '3',
        '四': '4',
        '五': '5',
```

```
        '六': '6',
        '七': '7',
        '八': '8',
    }
    cnt = {
        '1': 1,
        '2': 1,
        '3': 1,
        '4': 1,
        '5': 1,
        '6': 1,
        '7': 1,
        '8': 1,
    }
    for i in data:
        try:
            i['id'] = reflect[i['batch'][1]]+'-'+str(cnt[reflect[i['batch']
[1]]])
            cnt[reflect[i['batch'][1]]]+=1
        except:
            # 阿尔寨石窟
            # 焦裕禄烈士墓
            # 这两个属于第五批增补单位，未指定id，故单独标识
            print("error",i['name'])
    with open('result.json', 'w', encoding='utf-8') as f:
        json.dump(data, f, ensure_ascii=False, indent=4)

def location():
    with open('result.json', 'r', encoding='utf-8') as f:
        data = json.load(f)
    for i in range(4000,len(data)):
        query = data[i]['address']+data[i]['name']
        response = requests.get(
            url='https://restapi.amap.com/v3/geocode/geo',
            params={
                'key':'07519ede4916771d7c44b6f3bc3b0bfd',
                'address':query
            },
            headers=headers
        )
        try:
            print(i,data[i]['name'],response.json()['geocodes'][0]['location'])
            data[i]['location'] = response.json()['geocodes'][0]['location']
        except:
            print(data[i]['name'])
            with open('error.txt', 'a', encoding='utf-8') as f:
                f.write(data[i]['name']+'\n'+response.text+'\n\n')
        time.sleep(0.5)
        if i %100 == 0:
            with open('result.json', 'w', encoding='utf-8') as f:
                json.dump(data, f, ensure_ascii=False, indent=4)
    with open('result.json', 'w', encoding='utf-8') as f:
        json.dump(data, f, ensure_ascii=False, indent=4)

pre_process()
```

```
location()
```

在调用API时，有时会遇到

> {"status":"0","info":"ENGINE_RESPONSE_DATA_ERROR","infocode":"30001"}

的返回异常，所以引入`try...except...`纠错机制，对出现问题的数据单独保存，后续再重新调用接口，直至获取到数据为止，最终得到的数据包含字段如下：

```
{
    "id": "1-1",
    "name": "三元里平英团遗址",
    "batch": "第一批",
    "type": "革命遗址及革命纪念建筑物",
    "address": "广东省广州市三元里",
    "publish": "文物保护单位",
    "age": "1841年",
    "location": "113.267187,23.162749"
}
```