图像采集

导入工具包

In [1]:
```
import os

import time


import requests

import urllib3
urllib3.disable_warnings()

# 进度条库
from tqdm import tqdm

import os
```

HTTP请求参数 从百度上爬取图片

In [6]:
```
cookies = {
'BDqhfp': '%E7%8B%97%E7%8B%97%26%26NaN-1undefined%26%2618880%26%2621',
'BIDUPSID': '06338E0BE23C6ADB52165ACEB972355B',
'PSTM': '1646905430',
'BAIDUID': '104BD58A7C408DABABCAC9E0A1B184B4:FG=1',
'BDORZ': 'B490B5EBF6F3CD402E515D22BCDA1598',
'H_PS_PSSID': '35836_35105_31254_36024_36005_34584_36142_36120_36032_35993_35984_353
'BDSFRCVID': '8--OJexroGOxMovDbuOS5T78igKKHJQTDYLtOwXPsp3LGJLVgaSTEGOPtjcEHMA-2Zlgog
'H_BDCLCKID_SF': 'tJPqoKtbtDI3fP36qR3KhPt8Kpby2D62aKDs2nopBhcqEIL4QTQM5p5yQ2c7LUvtyn
'BDSFRCVID_BFESS': '8--OJexroGOxMovDbuOS5T78igKKHJQTDYLtOwXPsp3LGJLVgaSTEGOPtjcEHMA-
'H_BDCLCKID_SF_BFESS': 'tJPqoKtbtDI3fP36qR3KhPt8Kpby2D62aKDs2nopBhcqEIL4QTQM5p5yQ2c7
'indexPageSugList': '%5B%22%E7%8B%97%E7%8B%97%22%5D',
'cleanHistoryStatus': '0',
'BAIDUID_BFESS': '104BD58A7C408DABABCAC9E0A1B184B4:FG=1',
'BDRCVFR[dG2JNJb_ajR]': 'mk3SLVN4HKm',
'BDRCVFR[-pGxjrCMryR]': 'mk3SLVN4HKm',
'ab_sr': '1.0.1_Y2YxZDkwMWZkMmY2MzA4MGU0OTNhMzVlNTcwMmM2MWE4YWU4OTc1ZjZmZDM2N2RjYmVk
'delPer': '0',
'PSINO': '2',
'BA_HECTOR': '8h24a024042g05alup1h3g0aq0q',
}

headers = {
'Connection': 'keep-alive',
'sec-ch-ua': '" Not;A Brand";v="99", "Google Chrome";v="97", "Chromium";v="97"',
'Accept': 'text/plain, */*; q=0.01',
'X-Requested-With': 'XMLHttpRequest',
'sec-ch-ua-mobile': '?0',
'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (K
'sec-ch-ua-platform': '"macOS"',
'Sec-Fetch-Site': 'same-origin',
'Sec-Fetch-Mode': 'cors',
'Sec-Fetch-Dest': 'empty',
'Referer': 'https://image.baidu.com/search/index?tn=baiduimage&ipn=r&ct=201326592&cl
'Accept-Language': 'zh-CN,zh;q=0.9',
}
```

指定关键词

In [10]:
```
# 关键词
keyword = '香蕉'

# 拟爬取图像个数
DOWNLOAD_NUM = 5
```

创建文件夹

In [11]:
```
if not os.path.exists('dataset'):
    os.makedirs('dataset')
    print('新建 dataset 文件夹')

if os.path.exists('dataset/'+keyword):
    print('文件夹 dataset/{} 已存在，之后直接将爬取到的图片保存至该文件夹中'.format(
else:
    os.makedirs('dataset/{}'.format(keyword))
    print('新建文件夹：dataset/{}'.format(keyword))
```

新建文件夹：dataset/香蕉

爬取并保存图像文件至本地

In [12]:
```
count = 1

# 爬取第几张
num = 1

# 是否继续爬取
FLAG = True

while FLAG:

    page = 30 * count

    params = (
        ('tn', 'resultjson_com'),
        ('logid', '12508239107856075440'),
        ('ipn', 'rj'),
        ('ct', '201326592'),
        ('is', ''),
        ('fp', 'result'),
        ('fr', ''),
        ('word', f'{keyword}'),
        ('queryWord', f'{keyword}'),
        ('cl', '2'),
        ('lm', '-1'),
        ('ie', 'utf-8'),
        ('oe', 'utf-8'),
        ('adpicid', ''),
        ('st', '-1'),
        ('z', ''),
        ('ic', ''),
        ('hd', ''),
        ('latest', ''),
        ('copyright', ''),
        ('s', ''),
        ('se', ''),
        ('tab', ''),
        ('width', ''),
        ('height', ''),
        ('face', '0'),
        ('istype', '2'),
        ('qc', ''),
        ('nc', '1'),
        ('expermode', ''),
        ('nojc', ''),
        ('isAsync', ''),
        ('pn', f'{page}'),
        ('rn', '30'),
        ('gsm', '1e'),
        ('1647838001666', ''),
    )

    response = requests.get('https://image.baidu.com/search/acjson', headers=headers
    if response.status_code == 200: #状态数字200 正常 404未找到
        try:
            json_data = response.json().get("data")  #转换成json文件 解析json数据
            # print(json_data)
            if json_data:
                for x in json_data:
                    type = x.get("type")
                    if type not in ["gif"]:
                        img = x.get("thumbURL")  # 图片地址
                        fromPageTitleEnc = x.get("fromPageTitleEnc")  # 获取图片的标
```

```
                        try:
                            resp = requests.get(url=img, verify=False)
                            time.sleep(1)
                            print(f"链接 {img}")
                            # 保存文件名
                            # file_save_path = f'dataset/{keyword}/{num}-{fromPageTi
                            file_save_path = f'dataset/{keyword}/{num}.{type}'
                            with open(file_save_path, 'wb') as f:
                                f.write(resp.content) # 这个是直接从网络上面抓取的数
                                f.flush() #直接将缓冲区内容写入磁盘
                                print('第 {} 张图像 {} 爬取完成'.format(num, fromPag
                                num += 1


                            # 爬取数量达到要求
                            if num > DOWNLOAD_NUM:
                                FLAG = False
                                print('{} 张图像爬取完毕'.format(num))
                                break

                        except Exception:
                            pass
                except:
                    pass
            else:
                break

            count += 1
```

```
链接 https://img0.baidu.com/it/u=3130349898,4086493786&fm=253&fmt=auto&app=138&f=JPE
G?w=647&h=500
第 1 张图像 新鲜的香蕉照片-正版商用图片1e42ed-摄图新视界 爬取完成
链接 https://img0.baidu.com/it/u=1725638582,583115983&fm=253&fmt=auto&app=138&f=JPG?
w=500&h=730
第 2 张图像 香蕉元素素材下载-正版素材401753335-摄图网 爬取完成
链接 https://img1.baidu.com/it/u=4039888881,3323377242&fm=253&fmt=auto&app=138&f=JPE
G?w=500&h=667
第 3 张图像 香蕉-堆糖,美好生活研究所 爬取完成
链接 https://img0.baidu.com/it/u=4050209701,3599579398&fm=253&fmt=auto&app=138&f=JPE
G?w=500&h=334
第 4 张图像 一根香蕉防治八种病_中华康网 爬取完成
链接 https://img0.baidu.com/it/u=2443303341,2850529792&fm=253&fmt=auto&app=138&f=JPE
G?w=588&h=438
第 5 张图像 真相：空腹不能吃香蕉吗?-刘萍萍-爱问医生 爬取完成
6 张图像爬取完毕
```

封装函数

In [6]:

```
def craw_single_class(keyword, DOWNLOAD_NUM = 200):
    if os.path.exists('dataset/'+keyword):
        print('文件夹 dataset/{} 已存在，之后直接将爬取到的图片保存至该文件夹中'.for
    else:
        os.makedirs('dataset/{}'.format(keyword))
        print('新建文件夹：dataset/{}'.format(keyword))
    count = 1

    with tqdm(total=DOWNLOAD_NUM, position=0, leave=True) as pbar:

        # 爬取第几张
        num = 0

        # 是否继续爬取
        FLAG = True
```

```
    while FLAG:

        page = 30 * count

        params = (
            ('tn', 'resultjson_com'),
            ('logid', '12508239107856075440'),
            ('ipn', 'rj'),
            ('ct', '201326592'),
            ('is', ''),
            ('fp', 'result'),
            ('fr', ''),
            ('word', f'{keyword}'),
            ('queryWord', f'{keyword}'),
            ('cl', '2'),
            ('lm', '-1'),
            ('ie', 'utf-8'),
            ('oe', 'utf-8'),
            ('adpicid', ''),
            ('st', '-1'),
            ('z', ''),
            ('ic', ''),
            ('hd', ''),
            ('latest', ''),
            ('copyright', ''),
            ('s', ''),
            ('se', ''),
            ('tab', ''),
            ('width', ''),
            ('height', ''),
            ('face', '0'),
            ('istype', '2'),
            ('qc', ''),
            ('nc', '1'),
            ('expermode', ''),
            ('nojc', ''),
            ('isAsync', ''),
            ('pn', f'{page}'),
            ('rn', '30'),
            ('gsm', '1e'),
            ('1647838001666', ''),
        )

        response = requests.get('https://image.baidu.com/search/acjson', headers
        if response.status_code == 200: # 404未找到
            try:
                json_data = response.json().get("data")
                # print(json_data)

                if json_data:
                    for x in json_data:
                        type = x.get("type")
                        if type not in ["gif"]:
                            img = x.get("thumbURL") #图片地址
                            fromPageTitleEnc = x.get("fromPageTitleEnc") #获取标
                            try:
                                resp = requests.get(url=img, verify=False)
                                time.sleep(1)
                                # print(f"链接 {img}")

                                # 保存文件名
                                # file_save_path = f'dataset/{keyword}/{num}-{fr
                                file_save_path = f'dataset/{keyword}/{num}.{type
```

```
                                with open(file_save_path, 'wb') as f:
                                    f.write(resp.content) # 抓取图片数据，保存图
                                    f.flush()  #直接将缓冲区内容写入磁盘
                                    # print('第 {} 张图像 {} 爬取完成'.format(nu
                                    num += 1
                                    pbar.update(1) # 进度条更新

                                # 爬取数量达到要求
                                if num > DOWNLOAD_NUM:
                                    FLAG = False
                                    print('{} 张图像爬取完毕'.format(num))
                                    break

                            except Exception:
                                pass
                    except:
                        pass
                else:
                    break

                count += 1
```

In [7]:
```
craw_single_class('柚子', DOWNLOAD_NUM = 2)
```

```
文件夹 dataset/柚子 已存在，之后直接将爬取到的图片保存至该文件夹中
3it [00:03,  1.30s/it]
3 张图像爬取完毕
```

In [8]:
```
class_list = ['黄瓜','南瓜','冬瓜','木瓜','苦瓜','丝瓜','窝瓜','甜瓜','香瓜','白兰瓜
```

In [10]:
```
for each in class_list:
    craw_single_class(each, DOWNLOAD_NUM = 200)
```

```
新建文件夹：dataset/黄瓜
2it [00:02,  1.29s/it]
2 张图像爬取完毕
新建文件夹：dataset/南瓜
2it [00:02,  1.36s/it]
2 张图像爬取完毕
新建文件夹：dataset/冬瓜
2it [00:02,  1.35s/it]
2 张图像爬取完毕
新建文件夹：dataset/木瓜
2it [00:02,  1.50s/it]
2 张图像爬取完毕
新建文件夹：dataset/苦瓜
2it [00:02,  1.38s/it]
2 张图像爬取完毕
新建文件夹：dataset/丝瓜
2it [00:03,  1.51s/it]
2 张图像爬取完毕
新建文件夹：dataset/窝瓜
2it [00:03,  1.56s/it]
2 张图像爬取完毕
新建文件夹：dataset/甜瓜
2it [00:02,  1.30s/it]
2 张图像爬取完毕
新建文件夹：dataset/香瓜
2it [00:02,  1.32s/it]
```

```
2 张图像爬取完毕
新建文件夹：dataset/白兰瓜
2it [00:02,  1.26s/it]
2 张图像爬取完毕
新建文件夹：dataset/黄金瓜
2it [00:02,  1.27s/it]
2 张图像爬取完毕
新建文件夹：dataset/西葫芦
2it [00:02,  1.34s/it]
2 张图像爬取完毕
新建文件夹：dataset/人参果
2it [00:02,  1.36s/it]
2 张图像爬取完毕
新建文件夹：dataset/羊角蜜
2it [00:02,  1.31s/it]
2 张图像爬取完毕
新建文件夹：dataset/佛手瓜
2it [00:02,  1.29s/it]
2 张图像爬取完毕
新建文件夹：dataset/伊丽莎白瓜
2it [00:02,  1.26s/it]
2 张图像爬取完毕
```