import requests  
  
headers = {  
 'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_12\_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/537.36',  
 'Referer': 'https://www.google.com/'}  
url = "https://movie.douban.com/top250"  
  
try:  
 response = requests.get(url, headers=headers)  
  
 print(f"网页文本:{response.text}")  
 print(f"返回网页文本的字节串: {response.content}")  
 print(f"响应状态码:{response.status\_code}")  
 print(f"响应头:{response.headers}")  
 print(f"请求头: {response.request.headers} ")  
  
 response.raise\_for\_status()  
 ss = response.content.decode()  
  
except Exception as e:  
 print("失敗",e)

下載圖片

import requests  
  
user\_agent = "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Geoko) Chrome/104.0.5112.102 Safari/537.36 Edg/104.0.1293.70"  
  
response = requests.get("https://img1.doubanio.com/view/subject/l/public/s34086258.jpg",  
headers= {'User-Agent': user\_agent})  
  
print(response)  
with open("web.png","wb")as f:  
 f.write(response.content)

5-2

import json  
import time  
from bs4 import BeautifulSoup  
import requests  
  
url = 'https://movie.douban.com/top250'  
user\_agent = 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/103.0.5060.114 Safari/537.36 Edg/103.0.1264.49'  
cookie = 'll="118147"; bid=epfLHDtMjFY; douban-fav-remind=1; \_\_gads=ID=e0a8a10ac8562492-22fa21ad2ad600df:T=1662186388:RT=1662186388:S=ALNI\_Mb8BzconcZNqcMqhR\_ao-TD88JtxA; \_\_gpi=UID=0000096b769c01ae:T=1662186388:RT=1662186388:S=ALNI\_MZuWlYt-vEQyWkoIwPr-RptBUUBLQ; ct=y; \_pk\_ref.100001.8cb4=["","",1662196853,"https://cn.bing.com/"]; \_pk\_ses.100001.8cb4=\*; \_\_utmz=30149280.1662196854.6.3.utmcsr=cn.bing.com|utmccn=(referral)|utmcmd=referral|utmcct=/; \_\_utmc=30149280; \_\_utma=30149280.1610394692.1651277974.1662192922.1662196854.6; \_\_utmt=1; dbcl2="262213071:mNUuOGkM5Iw"; ck=mpJD; push\_noty\_num=0; push\_doumail\_num=0; \_pk\_id.100001.8cb4=fd90b18bbe540a04.1644059397.5.1662197142.1662105755.; \_\_utmv=30149280.26221; \_\_utmb=30149280.4.10.1662196854'  
# user\_agent='Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:104.0) Gecko/20100101 Firefox/104.0'  
# cookie="bid=vY5I22IBQoM; \_\_utma=30149280.730987918.1654349972.1654349972.1654478157.2; \_\_utmz=30149280.1654478157.2.2.utmcsr=baidu|utmccn=(organic)|utmcmd=organic; \_\_gads=ID=5c6881f77ebf7f97-2224ed90c5d300b0:T=1654349955:RT=1654349955:S=ALNI\_MZwkY\_YcucUDfu2a5XZnI8-dEzW5Q; \_\_gpi=UID=000006489dc2fefe:T=1654349955:RT=1654478138:S=ALNI\_MYsAFDkRiB\_uX8FJCyfrZq-0wKA\_w"  
list = []  
  
while True:  
 response = requests.get(url, headers={'User-Agent': user\_agent, 'Cookie': cookie})  
 print(response)  
 soup = BeautifulSoup(response.text, 'html.parser')  
 results = soup.select('#wrapper #content .grid\_view li')  
  
 for item in results:  
 title = item.select('.info .hd span.title')[0].text  
 print(title)  
 director = \  
 item.select('.info .bd p')[0].text.strip().split('主演:')[0].strip().split('主...')[0].strip().split(':')[  
 1].strip()  
 print(director)  
 year = item.select('.info .bd p')[0].text.strip().split('\n')[1].strip().split('/')[0].strip()  
 print(year)  
 country = item.select('.info .bd p')[0].text.strip().split('\n')[1].strip().split('/')[1].strip()  
 print(country)  
 style = item.select('.info .bd p')[0].text.strip().split('\n')[1].strip().split('/')[2].strip()  
 print(style)  
 scores = item.select('.info .bd .star span')[1].text  
 print(scores)  
 number = item.select('.info .bd .star span')[3].text  
 print(number)  
 a = item.select('.item .hd a')[0]['href']  
 print(a)  
 response1 = requests.get(a, headers={'User-Agent': user\_agent, 'Cookie': cookie})  
 soup1 = BeautifulSoup(response1.text, 'html.parser')  
 content = soup1.select('#content .article [class="all hidden"]')  
 if len(content) != 0:  
 content = soup1.select('#content .article [class="all hidden"]')[0].text.strip()  
 content = ''.join(content.split())  
 else:  
 content = soup1.select('#content .article [property="v:summary"]')[0].text.strip()  
 content = ''.join(content.split())  
 print(content)  
 print('\*' \* 20)  
 dic = {'title': title, 'director': director, 'year': year, 'scores': scores, 'number': number, 'style': style,  
 'country': country, 'web': a, 'content': content}  
 list.append(dic)  
 time.sleep(2.0)  
  
 next\_url = soup.select('span.next a')  
 if len(next\_url) == 0:  
 break  
 url = 'https://movie.douban.com/top250' + next\_url[0]['href']  
 print(url)  
 time.sleep(5.0)  
  
with open('doubanTop250.json', 'w') as fp:  
 json.dump(list, fp)