這個python主要來抓去豆瓣(大陸最出名的影評網站)網上的數據，

源代碼：# -\*- coding: cp936 -\*-

import urllib.request

import re

import os

import time

import xlwt

def movie(movieTag):

tagUrl=urllib.request.urlopen(url)

tagUrl\_read = tagUrl.read().decode('utf-8')

return tagUrl\_read

def subject(tagUrl\_read):

#正则表达式匹配电影的名字（链接）、评分与评论

nameURL= re.findall(r'\s+title="(.+)"',tagUrl\_read)

scoreURL = re.findall(r'<span\s+class="rating\_nums">([0-9.]+)<\/span>',tagUrl\_read)

evaluateURL = re.findall(r'<span\s+class="pl">\((\w+)人评价\)<\/span>',tagUrl\_read)

movieLists = list(zip(nameURL,scoreURL,evaluateURL))

newlist.extend(movieLists)

return newlist

def find\_imgs(url):

html = movie(url)

img\_addrs = []

a = html.find('img src=')

while a != -1:

b = html.find('.jpg', a, a+255)

if b != -1:

img\_addrs.append(html[a+9:b+4])

else:

b = a + 9

a = html.find('img src=', b)

return img\_addrs

def save\_imgs(img\_addrs,no):

for each in img\_addrs:

filename=str(no)

with open(filename, 'wb') as f:

response = urllib.request.urlopen(each)

img = response.read()

f.write(img)

no+=1

#用quote处理特殊（中文）字符

movie\_type = urllib.request.quote(input('请输入电影类型(如剧情、喜剧、悬疑)：'))

page\_end=int(input('请输入搜索结束时的页码：'))

num\_end=page\_end\*20

num=0

page\_num=1

count=1

no=1

newlist=[]

os.mkdir(movie\_type)

os.chdir(movie\_type)

while num<num\_end:

url=r'http://movie.douban.com/tag/%s?start=%d'%(movie\_type,num)

movie\_url = movie(url)

subject\_url=subject(movie\_url)

img\_addrs = find\_imgs(movie\_url)

save\_imgs(img\_addrs,no)

num=page\_num\*20

page\_num+=1

else:

#使用sorted函数对列表进行排列，reverse参数为True时升序，默认或False时为降序， key=lambda还不是很明白这里的原理

movieLIST = sorted(newlist, key=lambda movieList : movieList[1],reverse = True)

f = open('demo.txt','w')

file=xlwt.Workbook()

table=file.add\_sheet('data')

table.write(0,0,'name')

table.write(0,1,'score')

table.write(0,2,'evaluate')

for i in range(len(movieLIST)):

m=i+1

table.write(m,0,movieLIST[i][0])

table.write(m,1,movieLIST[i][1])

table.write(m,2,movieLIST[i][2])

table = file.add\_sheet('sheet\_name',cell\_overwrite\_ok=True)

file.save('demo.xls') # 保存文件

# 另外，使用style

style = xlwt.XFStyle() # 初始化样式

font = xlwt.Font() # 为样式创建字体

font.name = 'Times New Roman'

font.bold = True

style.font = font #为样式设置字体

table.write(0, 0, 'some bold Times text', style) # 使用样式

for movie in movieLIST:

k=str(count)+":"+str(movie)

f.write(k+"\n")

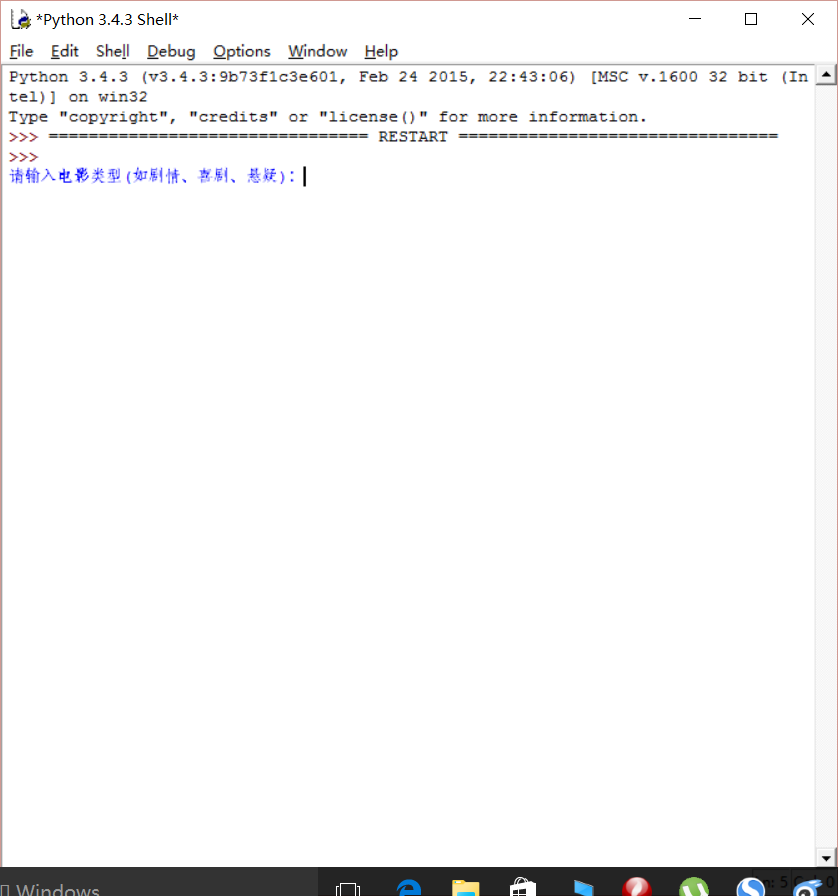
print(movie)

count+=1

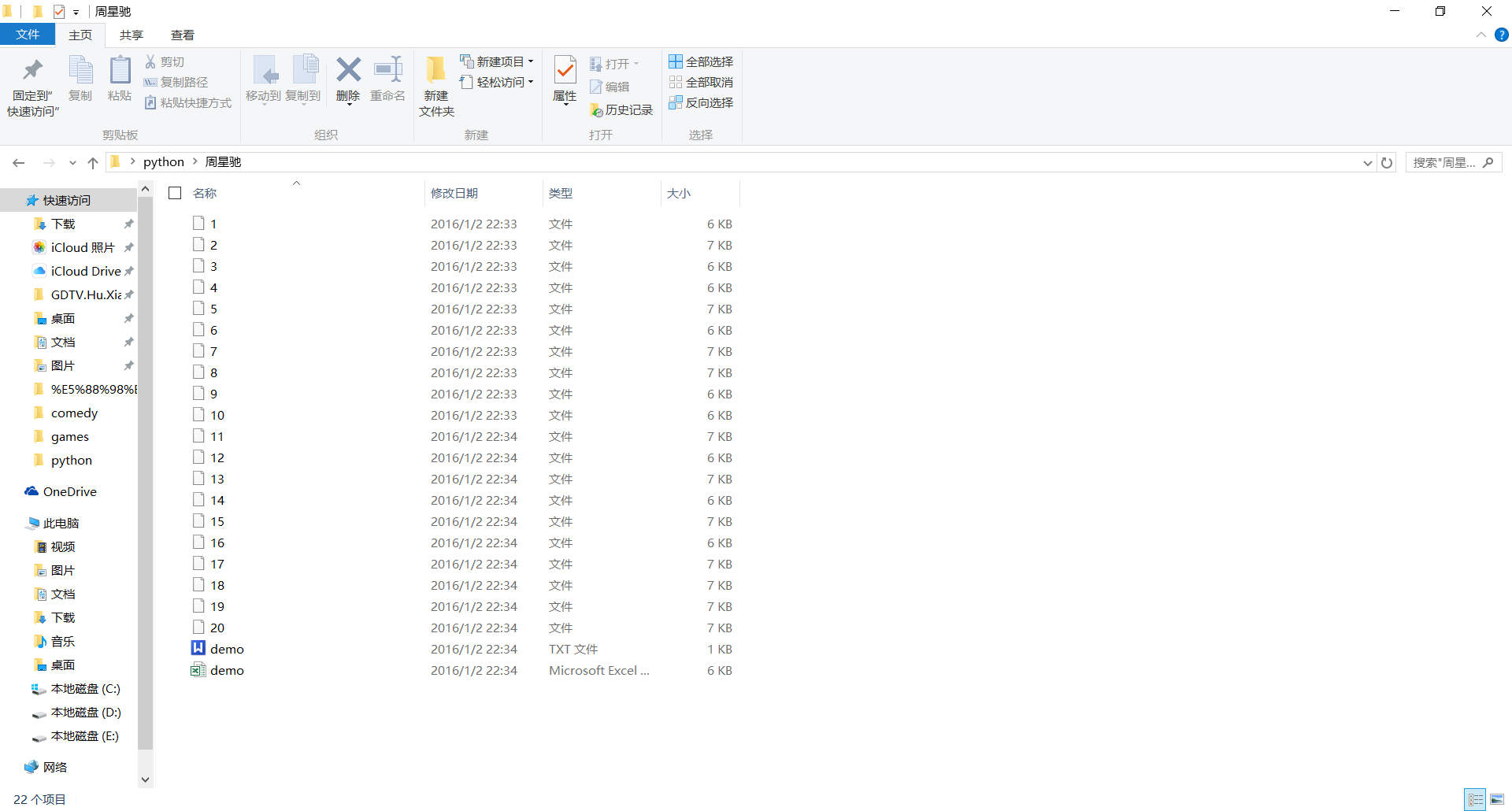
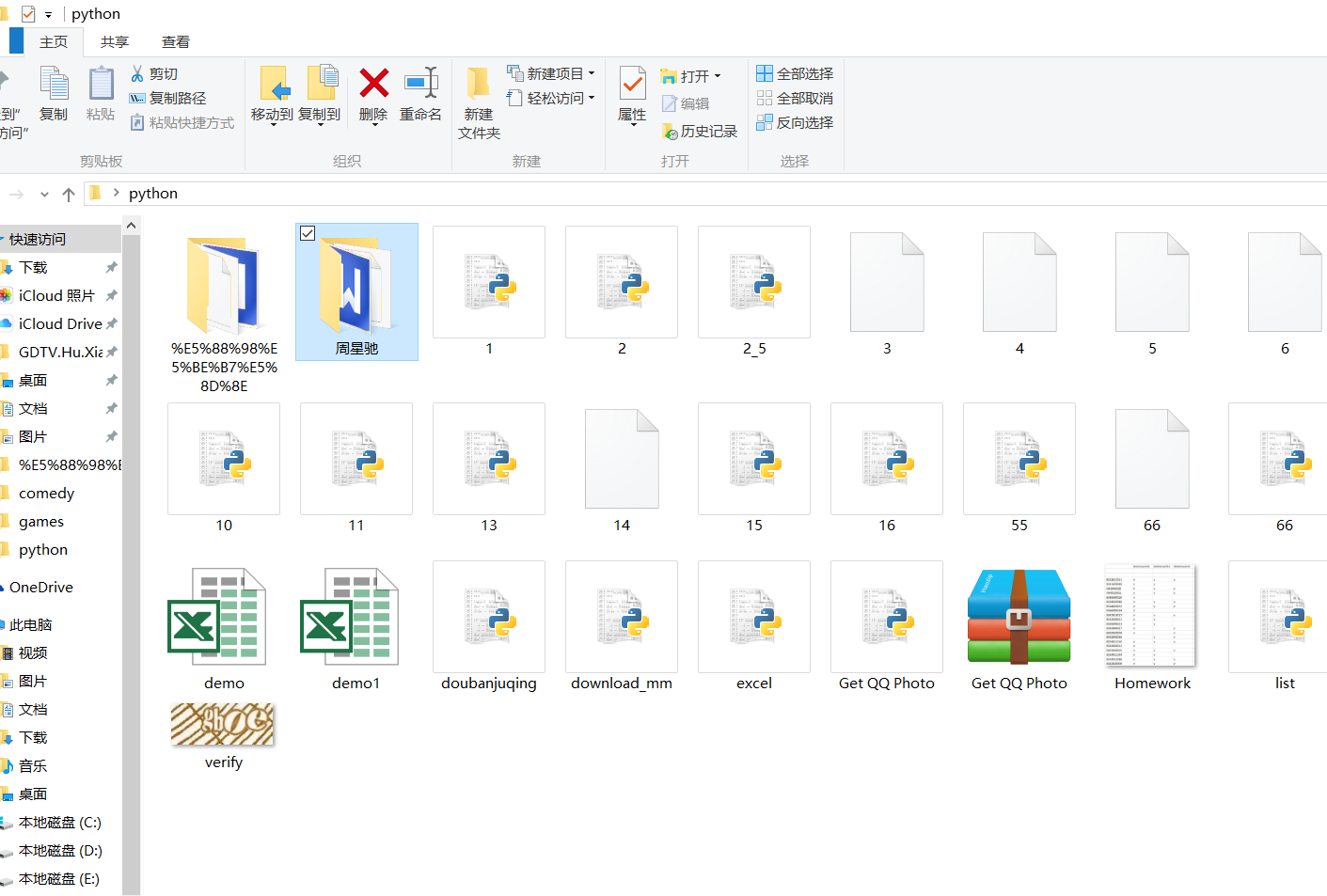
f.close()

time.sleep(3)

print('结束')



同時生成一個demo.txt檔將抓取的數據存儲在裏面。和对应的电影封面（我用find函数，





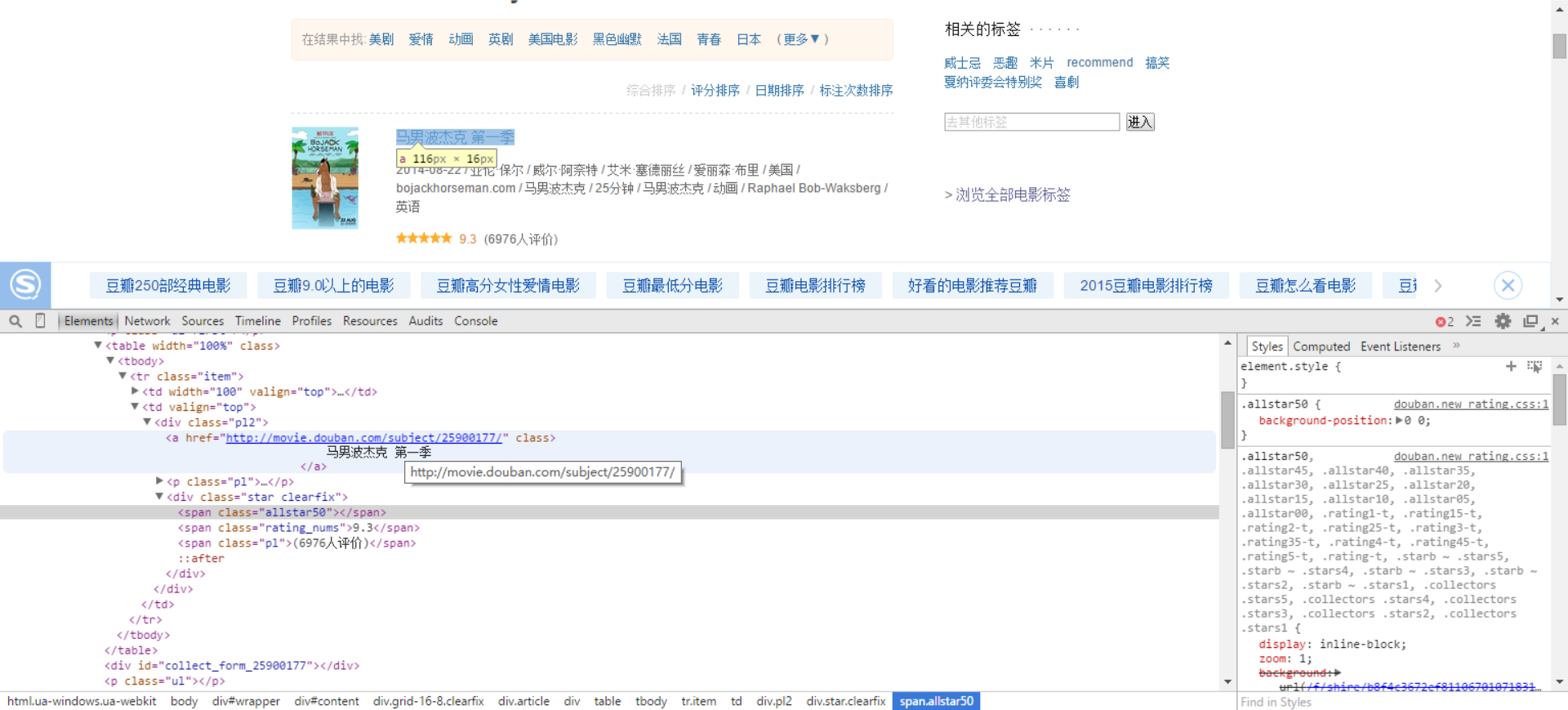
寫代碼的過程：進入豆瓣網站。

當我要搜索我想要看的電影影評：如：input：comedy。

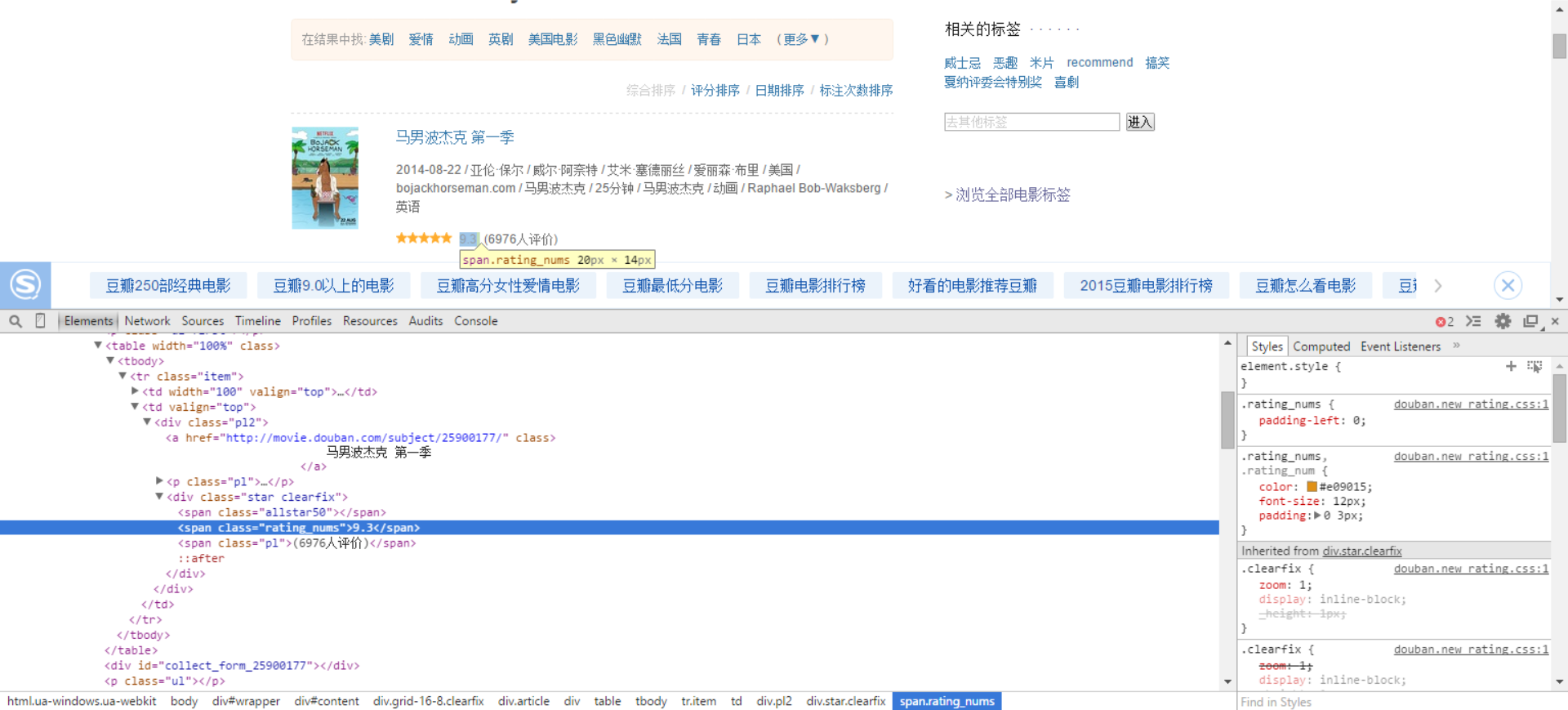
通過審查元素（F12）觀察dns變化，

http://movie.douban.com/tag/**comedy**?start=***0***

發現：tag後面就是我們想要搜索的類型，start後面就是page



Scores：



評價：

