网络爬虫

下载网页上的图片代码如下函数：imgUrl是图片的网址；destUrl是存储在本地的地址；fname是图片保存时的名字，默认值为空，图片保存时按网络上图片的名字保存

def downloadImageFile(imgUrl, destUrl, fname=''):

local\_filename = imgUrl.split('/')[-1]

print('Download Image File={}'.format(local\_filename))

try:

r = requests.get(imgUrl, stream=True)

r.raise\_for\_status()

if len(fname) == 0:

fname = local\_filename

print('fname={}'.format(fname))

with open(destUrl + "/" + fname, 'wb') as f:

for chunk in r.iter\_content(chunk\_size=1024):

if chunk:

f.write(chunk)

f.flush()

f.close()

return r.status\_code

except:

return r.status\_code

注意分析网页HTML数据，打开网页的开发者模式，可以看到图片的数据与来源

以爬取大学排名为例，获取排名信息时，有的网址做了反爬取，所以要在递交申请的循环中加入从网页中打开的步骤

如import webbrowser

IEPath = "C:\Program Files\internet explorer\iexplore.exe"

webbrowser.register('IE', None, webbrowser.BackgroundBrowser(IEPath))

webbrowser.get('IE').open('http://www.baidu.com', new=1,autoraise=True)

# 或者

webbrowser.open\_new\_tab('http://www.baidu.com')