# 学习笔记

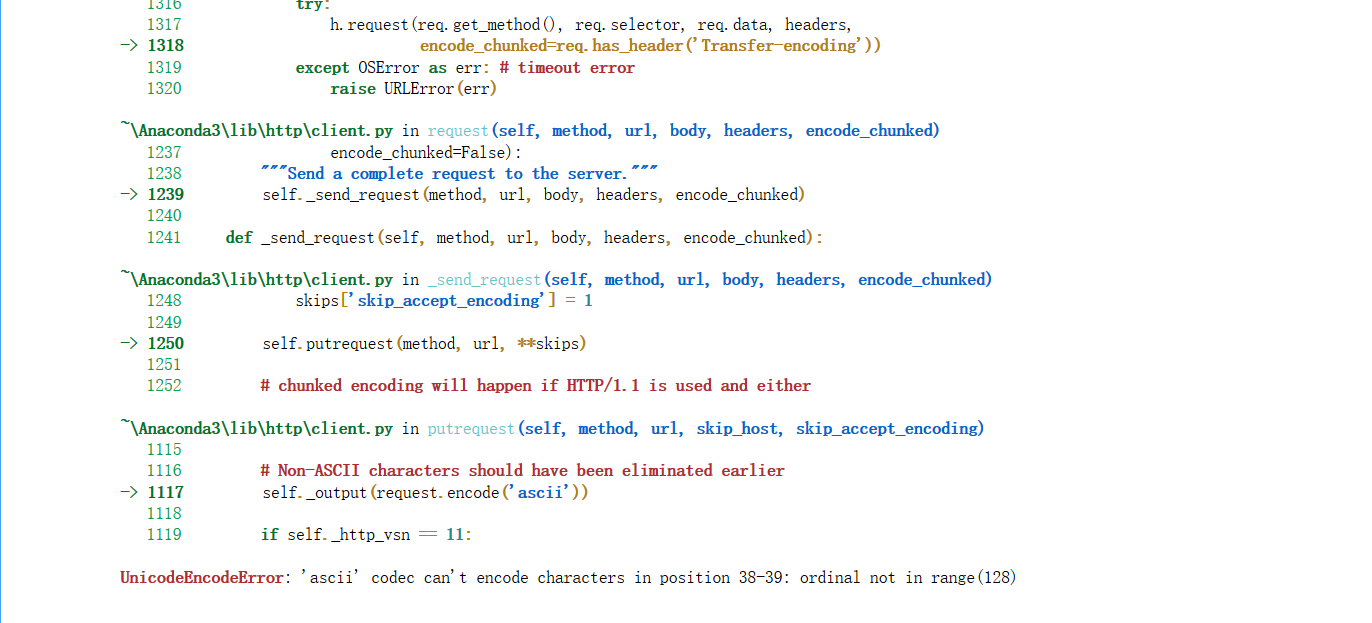
## 本周任务：

1. 实现了爬取豆瓣的电影信息（最终爬取了2000条左右）
2. 学习了一部分崔庆才编著的《python3网络爬虫开发实践》

## 遇到的问题：

1. **python3中url包含中文编码问题**

爬取电影时是根据获取到的豆瓣网的不同的电影分类标签来不断改变url实现获取的，但是豆瓣的标签获取到之后是包含中文的，在使用urlopen()来获取网站内容时，会出现如下错误：



通过查阅资料以及不断的尝试，根据python版本的不同，有两种解决方式，如果使用python2.7版本，可以在代码中插入如下代码，一次性解决编码问题：



如果是使用python3.x，由于在python3中setdefaultencoding()方法已经不存在了，所以需要使用其他方法来解决编码问题，尝试过将标签中的中文全部使用urlencoding编码一下，但是这种方式太过繁琐，最后发现可以使用python3中的quote来解决url中英文混编出现的编码问题，代码如下图所示：



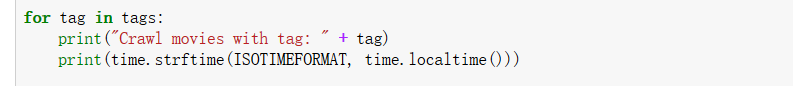
首先要导入quote，在python3.6x中需要注意的是，导入quote的方式是from urllib.parse import quote，即在urllib的parse模块内，直接从urllib中导入会出现找不到模块的问题，而url = quote(url, safe=string.printable)中，safe属性如此设定就是将url中可以输出的字符原样输出，不可以的（如中文），进行编码，插入这两行代码之后，再次运行就不会再出现之前的编码错误。

1. **关于时间格式**

在写代码过程中为了记录时间并输出，涉及到一些时间格式的设定和输出问题，在python中，也是先设定时间格式，然后通过函数将时间转换为指定格式的当地时间输出，设定时间格式的方式如下所示（可以根据实际将格式设定成其他形式）：

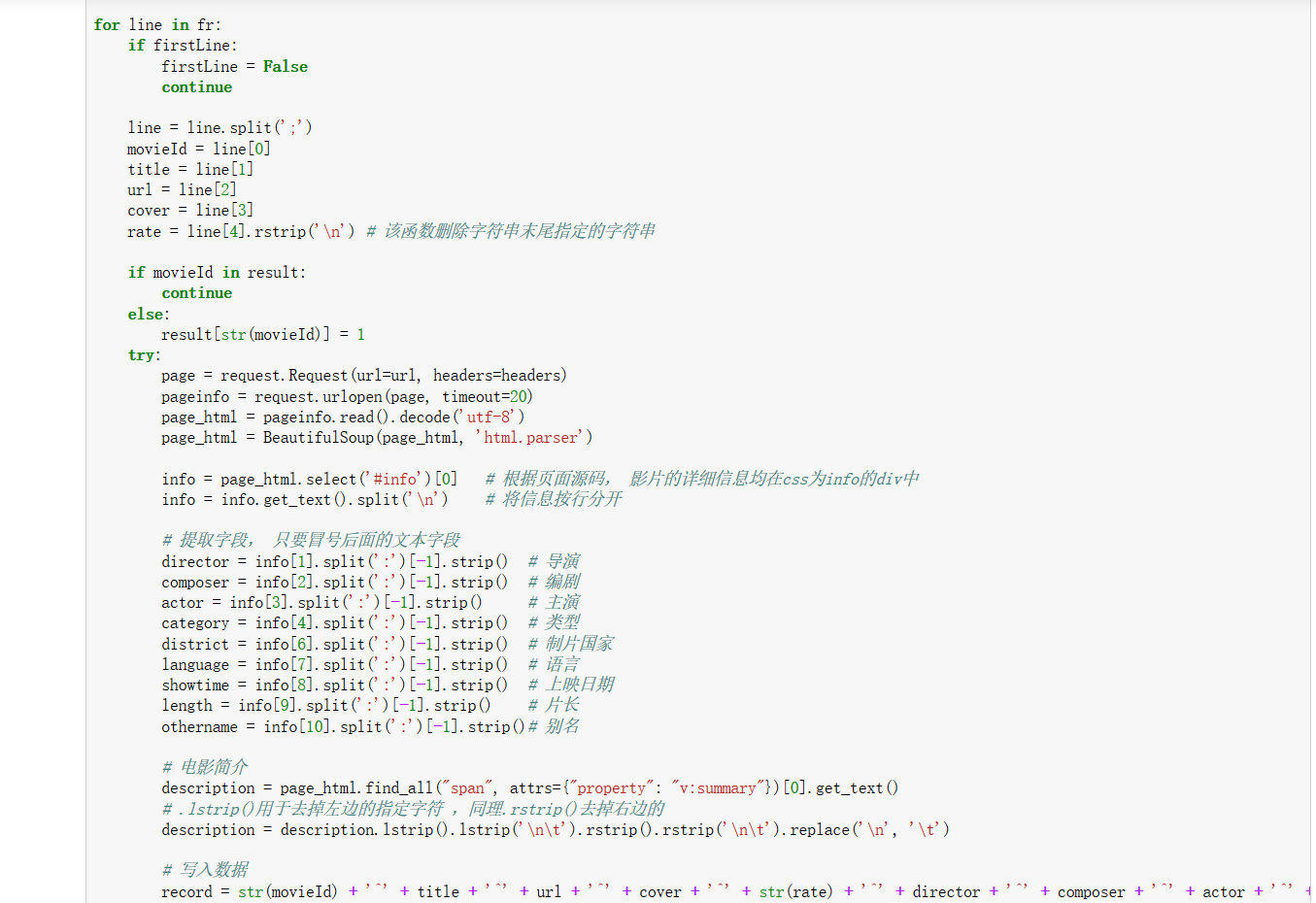


获取时间并转化为指定格式需要使用time模块中的函数time.strftime(agr1, arg2)，该函数接收以时间元组，并返回以可读字符串表示的当地时间，格式由参数刚才设定的格式决定，具体使用如下：



1. **获取电影的详细信息**

在获取电影的详细信息的过程中，在写入文本文件的过程中，同样遇到了编码问题，后来发现原因是在写入时自己对中文进行了utf-8形式的编码之后才写入，将编码过程删除即可，获取详细信息的主要代码如下：



在使用BeautifulSoup的时候，一定要指定一个网页的解析器，否则在解析网页的过程中会报警告，此处选用的是html.paser解析器，并且每获取到一条信息就暂停一个时间，防止豆瓣封禁。

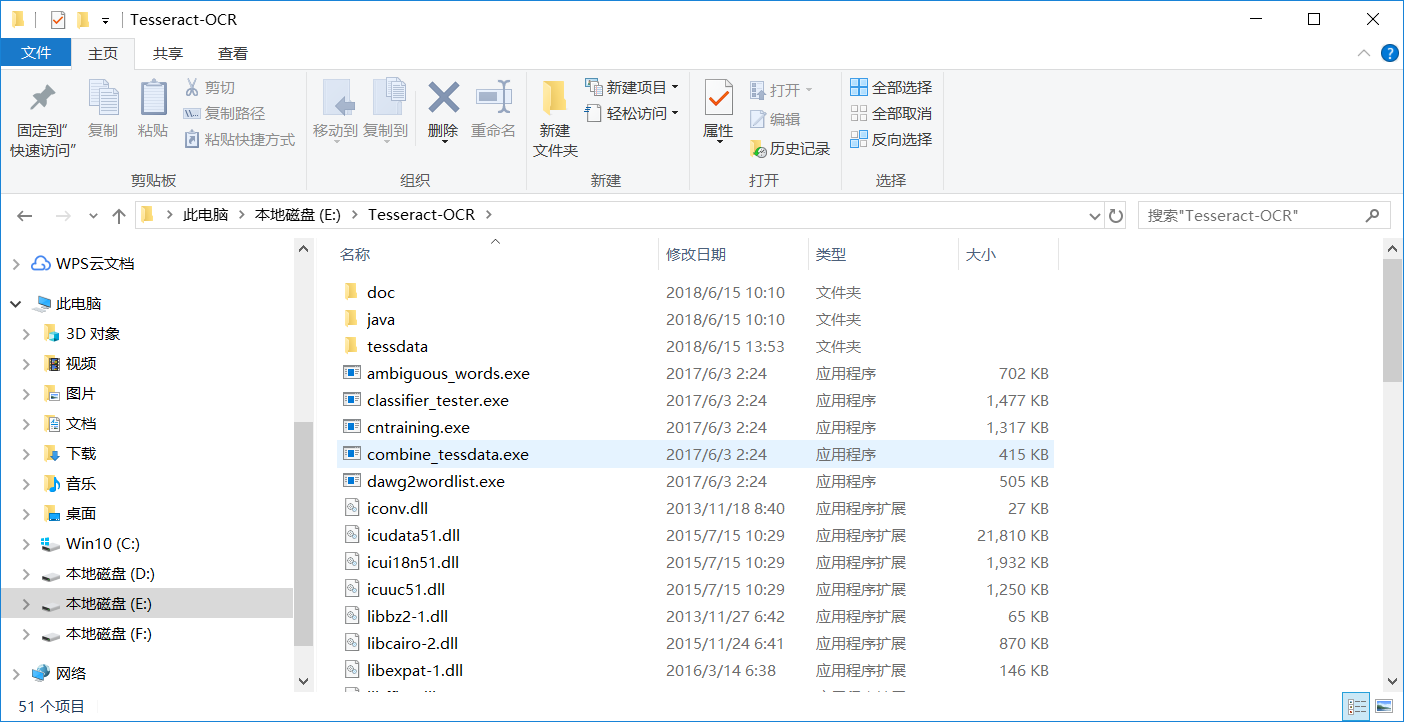
1. **尝试解决豆瓣的IP封禁**

在爬取过程中，最开始选择的sleep时间为5秒，但是爬取到六百多条的时候出现了403 Forbidden的错误，豆瓣拒绝了当前IP的访问，为了更换IP，将路由器重启，并将sleep时间改到了1秒，在爬取了1800条左右的时候再次被封禁，最后尝试了模拟登陆，网上查询模拟登陆的方式有两种，第一种是使用POST请求将账号密码发送到登陆表单的指定url，过程如下（会存在验证码问题）：

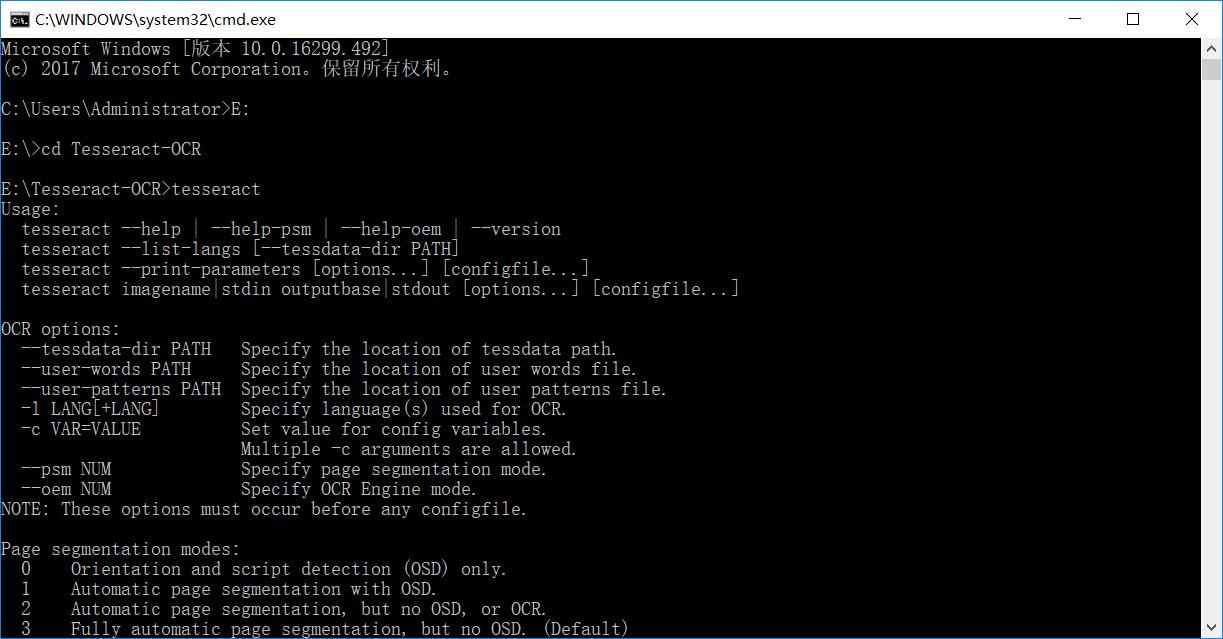


但是最后还是模拟登陆失败了，目前不知道如何解决，还有一种方式是通过设置cookie来模拟登陆，需要自己在豆瓣登陆后用火狐浏览器的开发者工具获取自己当前的cookie，将cookie的值设置在http的请求头中，实现模拟登陆，但是也没试成功，通过这种尝试最终使得我决定找一本比较系统的爬虫教材，实现绕过一些网站的反爬。

1. **关于《python3网络爬虫开发实践》的一些坑**
2. pyquery的安装：pyquery是一个强大的网页解析工具，他提供了和jQuery类似的语法来解析HTML文档，支持css选择器，但是在安装过程中出现了明明使用pip安装成功，但在尝试导入该模块的时候出来了没有该模块的错误提示，百度很久无法解决，后来通过咨询大佬，发现是由于自己安装了Anaconda的缘故，后来通过指令：conda install pyquery 成功安装并实现了导入。（对于先安装python后安装Anaconda的人来说，很容易出现包管理混乱的情况，遇到模块安装成功但导入失败的情况，大多可以通过这个方式解决，但最好是配置一下环境变量。。。。。目前我还没去实现）；
3. 关于tesseract的安装：
4. 安装OCR：OCR，即光学字符识别，是指通过扫描字符，然后通过其形状将其翻译成电子文本的过程，tesserocr是Python的一个OCR识别库，但其实是对tesseract做的一层Python API封装，故其核心是tesseract，在安装tesseract之前需要先安装tesseract，（切身体会：安装过程中选择自己需要的语言包即可，例如数学符号以及中文的，不需要全部都勾选，时间太长也用不到！），安装成功之后其是没有运行界面的，其安装目录结构如下所示：



由于没有运行界面，需要在cmd中运行，通过cd进入安装目录（不进入的话会提示不是内部或外部指令，或者需要配置环境变量），输入tesseract出现如下界面则证明安装成功：



1. 关于tesserocr和pillow的安装：在安装tesserocr的过程中可能会出现各种错误，例如缺少“setup”文件，缺少visual c++运行库，字符编码问题，此处强烈不推荐使用pip形式直接安装，可在网上找到whl格式的安装包，放在磁盘中 使用pip3 install tesserocr-2.2.2-cp36-cp36m-win\_amd64.whl命令来执行安装，但是，安装了Anaconda的同学用此方法安装成功后可能还是会出现找不到tesserocr模块无法导入的情况，此时需要执行如下指令：conda install -c simonflueckiger tesserocr 进行安装！！！！！

还有，如果在安装过程中出现了UnicodeDecodeError 这种编码错误，需要在Python的安装目录中修改一个东西，具体为在安装目录C:\Users\Administrator\AppData\Local\Programs\Python\Python36\Lib\site-packages\pip\compat中找到\_\_init\_\_.py文件，修改75行中的编码，将utf8改为gbk（或者其他形式），修改之后再执行安装就可以了，如下图所示：

