工作日记20171222静态页面抓取

Ringo Yungpo Kao

December 22, 2017

# 静态页面抓取

## 内容

本次主要看了抓取静态页面的基础知识，并实战抓取了TOP250电影数据，对于静态抓取页面有了进一步的人是。

## 基础知识

基础知识包括获取响应内容、定制Requests，也就是响应内容的获取，和请求参数的设置。 其中用requests.get放法来获取响应体。比如r = request.get(‘http://www.jmu.edu.cn’)。则可以通过r.text,r.encoding等查看r的参数来获取相应内容、文本编码、响应状态码等数据。 在使用request.get()方法的过程中，可以传入params传入请求参数，使用request.post时传入data参数，当然参数都是键值对的格式，所以要用字典的数据来传。

## TOP250电影数据爬取

爬取过程中重点要分析网站的特点，比如标题数据怎么查找，分页数据怎么提取，在足够地观察下就可以开始动手爬取了。本次爬取还算顺利，感觉理解起来还算好理解，不像scrapy那样难以理解。这次的抓取主要是用requests获取内容，然后用BeautifulSoup来进行分析。