

实验四-五报告

1. 安装Lucene

1.1. 原理简述

Lucene是一个全文检索引擎的架构。我们在爬取网页之后如果把网页内容都储存起来并且在搜索时采用遍历方法读取网页内容，无疑是非常耗用储存空间和搜索时间的。因此，我们需要对爬下来的内容进行分析，如上课讲的建立倒排索引等方法，建立并储存索引，也就不需要将原网页储存，并且搜索性能能大大提高。由于Lucene是基于Java环境的，安装之具有一定的复杂性。

2.2. Mac配置Lucene

Linux虚拟机在我的Mac上运行实在太慢，并且考虑到Mac是Unix系统，执行命令行也比较方便，所以我暂时没有使用虚拟机，项目都是在Mac环境下运行的。但由于助教给的教程都是基于Linux的，我在Mac上安装只能自行摸索。正好网上Mac安装Lucene的中文教程不是很丰富，以下是我Mac安装的步骤和遇到的各种问题，或许对他人有所帮助。

首先我考虑用教程中较简单的conda install方法。

①配置miniconda环境。在终端打开下载的conda.sh文件，常规安装等。但要注意，由于我的Mac系统本身有自带Python2，Pycharm也安装了Python3，这样系统中就有三个Python了，在终端调用时使用的是哪个需要注意。最好定位到路径。也可以调整默认路径，也就是修改~/.bashrc，设置环境变量，直至which python显示想要的路径。

②`conda install -c kalefranz pylucene=4.9.0`

按教程运行该命令后，conda会自动帮你把lucene与其前置库（包括jdk的静态库以及jcc）全部安装。但是，我每次都报错：找不到下载源文件。我第一反应是没有配置好channel，给conda的channel添加了国内的清华、豆瓣等镜像，但是多次尝试仍无用。最终我放弃了简单安装方法。

于是换用PPT中的第二种方法，一步步下载包和安装。

①安装ant

②`easy_install jcc`

③下载Pylucene，并编辑其Makefile文件，将其Python路径指向conda的Python。接下来运行make，按照PPT上内容，经过漫长的安装之后，最后不报错，在Python就可以调用lucene库了。然而，经过漫长的安装之后，我最终报错“jcc中的Pylucene不是Shared模式”。那么就去修改jcc的配置，但是在哪修改、怎么修改，查不到，只能自己摸索。在jcc整个文件夹内搜索’shared’字段，最终将python.py文件中一行Enable_Shared变量修改为True。再次运行，仍然报错。

④思考：报错显示运行时某权限错误。查资料，发现Pylucene安装时，需要先配置好jcc环境。而jcc的安装在Linux和Mac系统是不同的，在Mac中默认共享模式是False。现在报错是修改了Enable_Shared变量导致的。我使用easy_install会让包安装好，在安装过程中整个环境都配好了，最终也自动删除了安装包（包含setup.py文件），则很多不包含’shared’字段的内容也和Enable_Shared变量有关，只修改这一处显然不行。我们要使安装过程在shared==1的模式下运行，因此要修改安装前的命令。因此easy_install是不可靠的。

⑤分步进行。卸载先前安装的jcc，然后首先使用`pip download`下载jcc安装包，一个tar.gz文件。解压，其中果然有setup.py文件。直接install安装就是运行了该文件，并在运行完后删除。我修改了这个文件中的Enable_Shared变量为True，同时和一位有同样问题的同学讨论后，发现一个玄学变量with_modern_setuptools，会识别之前安装的某些库并导致莫名报错，于是将该变量整个改为True。

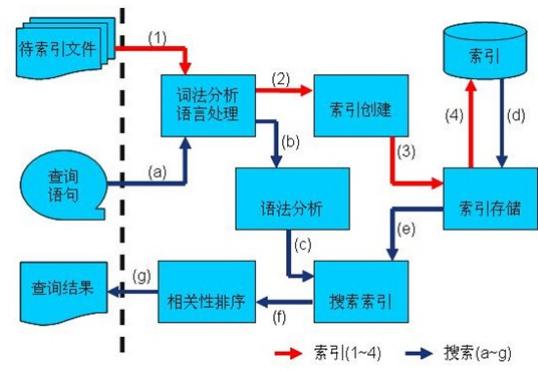
⑥修改完之后，cd进入jcc安装包，运行`sudo python setup.py install`。之后再执行③中修改Makefile等步骤，就安装好PyLucene在Python中了。`import lucene`，大功告成！

以上步骤看上去简单清晰，实际上耗用了整整一天时间。不是每个报错都能立即分析出原因并想出解决办法，需要耐心。这里感谢无私提供帮助的戚振林同学，指导了如何修改安装配置。

2.实现一个中文网页的索引与搜索程序

2.1.原理简述

原理部分在PPT中讲得十分清楚了。遍历爬虫爬下来的每一个网页，对其建立一个Document对象，该对象可以设置很多Field域，比如我设置了title（名称）、url、path（在本机储存位置）、filename等。这些域是只储存、不索引的。而另外设置的一个contents域，则是只索引而不储存的。contents来自结巴分词处理后的整个网页中的文字信息。最后将这个Document对象写入储存索引的文件夹中。而搜索程序则类似将这个过程反过来，用户键入的command，由结巴分词处理后，作为索引传入，并返回这些索引找到的文档。需要注意的是，command的预处理方式，必须与contents的预处理方式一致。



图示：索引+搜索

2.2.运行结果

2.2.1.查看索引

怎样查看自己的索引建立好了没有呢？我发现了一个和Lucene配套使用的可视化工具Luke。下载和自己Lucene版本一致的Luke，运行该java脚本，在弹出窗口中打开放置索引的文件夹，就可以看到索引的可视化信息了。

在Luke中查看索引和文档信息

2.2.2. 执行搜索

Searching for: 上海交通大学

title: 考研专区-权威考研专区分享资料平台_精品文库 url: <http://wenku.baidu.com/org/zone?zoneid=51> score: 0.402349174023 contents: None

title: 全国高校排名_全国大学排行榜_大学报考指南_新浪院校库_新浪教育 url: <http://edu.sina.com.cn/college/> score: 0.369252473116 contents: None

title: 海通证券 url: <https://www.hsec.com/ChannelHome/index.shtml> score: 0.301561534405 contents: None

title: 上海师范大学_百度百科 url: <https://baike.baidu.com/item/%E4%B8%A6%E6%85%87%E5%BB%88%E8%8C%83%E5%A4%A7%E5%AD%A6> score: 0.292811125517 contents: Nor

title: 大学排行榜实用指南：分层次报考_武书连_新浪博客 url: <http://blog.sina.com.cn/b/42bc00e0102xui.htm?tj=edu&tj=1> score: 0.284077644348 contents: I

title: 华东师范_王先庆_海上丝绸之路的内涵、形成与演变_王先庆_新浪博客 url: http://blog.sina.com.cn/s/blog_487ada310102yhtp.html score: 0.276518732309 contents: Nor

title: 王先庆：海上丝绸之路的内涵、形成与演变_王先庆_新浪博客 url: http://blog.sina.com.cn/s/blog_487ada310102yhtp.html score: 0.270677171998 contents: None

title: 上海（中华人民共和国直辖市）_百度百科 url: <https://baike.baidu.com/item/%E4%B8%A6%E6%85%87%E5%BB%88%E8%8C%83%E5%A4%A7%E5%AD%A6> score: 0.267726838589 contents: None

title: 高考分数线查询_全国各地各批次高考分数线查询_新浪教育 url: <http://kaoshi.edu.sina.com.cn/college/scorelist?year=2015> score: 0.266133368015 contents: N

title: 高考估分选大学_全国前10名大学_新浪院校库_新浪教育 url: <http://kaoshi.edu.sina.com.cn/college/collegelist/view?tab=contact> score: 0.263111442327 conte

title: 海德堡大学_百度百科 url: <https://baike.baidu.com/item/%E6%B5%87%E5%BB%88%E8%8C%83%E5%A4%A7%E5%AD%A6> score: 0.260530302525 contents: None

title: 海南师范大学_百度百科 url: <https://baike.baidu.com/item/%E9%9D%92%E6%B5%87%E5%BB%88%E8%8C%83%E5%A4%A7%E5%AD%A6> score: 0.250871866941 contents: Nor

title: 青海师范大学_百度百科 url: <https://baike.baidu.com/item/%E6%B5%87%E5%BB%88%E8%8C%83%E5%A4%A7%E5%AD%A6> score: 0.244668185711 contents: None

title: 同济大学_百度百科 url: <https://baike.baidu.com/item/%E5%99%8C%E6%B5%87%E5%BB%88%E8%8C%83%E5%A4%A7%E5%AD%A6> score: 0.243275418878 contents: None

title: 综合性大学_百度百科 url: <http://baike.baidu.com/view/1001725.html> score: 0.240761369467 contents: None

title: 龙虎榜 - 数据中心 - 新浪财经 url: <http://vip.stock.finance.sina.com.cn/q/go.php/vInvestConsult/kind/lhb/index.shtml> score: 0.238947331905 contents

title: 【新海外】出国留学_移民_海外房产_置业投资买房_中介机构口碑_专家咨询平台 url: <http://www.xhhouse.com/> score: 0.237220898271 contents: None

title: 格廷根大学_百度百科 url: <https://baike.baidu.com/item/%E5%8D%95%8C%5B%8B%87%E6%A0%89%5A%4A%7%E5%AD%A6> score: 0.235422223806 contents: None

title: 南通大学_百度百科 url: <https://baike.baidu.com/item/%E5%8D%97%E9%80%9A%5A%4A%7%E5%AD%A6> score: 0.235226795077 contents: None

title: 哥伦比亚大学_百度百科 url: <https://baike.baidu.com/item/%E7%BE%BE%5E%9B%BD%5E%93%5A%4A%BC%6A%F9%4A%6A%9A%5E%4A%7%E5%AD%A6> score: 0.22859223

title: 伊利诺伊大学厄巴纳-香槟分校_百度百科 url: <https://baike.baidu.com/item/%E4%BC%8A%5E%8B%8A%5E%8A%AF%BA%4E%BC%8A%5E%4A%7%E5%AD%A6%5E%8E%84%5E%87%BA%4E%7%5E> title: 中国客车网_客车行业门户网站_创立于1999年 url: <http://www.chinabuses.com/> score: 0.220905199647 contents: None

title: 南京大学_百度百科 url: <http://baike.baidu.com/view/3143.htm> score: 0.220905199647 contents: None

title: 华中师范大学_百度百科 url: <https://baike.baidu.com/item/%E5%8D%8E%4B%8B%AD%5E%8B%88%E8%8C%83%E5%A4%A7%E5%AD%A6> score: 0.22063575685 contents: None

title: 印度洋_百度百科 url: <https://baike.baidu.com/item/%E5%8D%80%5B%8A%6A%6B%84%88> score: 0.219284728169 contents: None

title: Title Not Found url: <http://www.szse.cn/> score: 0.2192455972 contents: None

title: 印第安纳大学_百度百科 url: <https://baike.baidu.com/item/%E5%8D%80%5B%8A%6A%6B%84%88> score: 0.216867342591 contents: Nor

title: 【上海二手车_上海二手车市场_上海二手车交易市场】_人人车 url: <https://www.renrenche.com/sh/> score: 0.214989542961 contents: None

title: 金网艺购商城_新浪网 url: http://collection.sina.com.cn/zt_d/jwvg/ score: 0.213803395629 contents: None

title: 北京师范大学_百度百科 url: <https://baike.baidu.com/item/%E5%8C%97%4B%8A%AC%5B%8B%88%E8%8C%83%E5%A4%A7%E5%AD%A6> score: 0.213526725769 contents: Nor

... total matching documents: 32321075629 contents: None

搜索“上海交通大学”

2.3. 细节问题

① 搜索出现重复结果。如下图所示：

拙/心/圆/曰/及/日/科 url: <http://baike.baidu.com/view/32201.htm> score: 0.180290045/8 contents: None

中国美术家协会 url: <http://www.caanet.org.cn/> score: 0.184433162212 contents: None

家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F%E9%9B%86%5E%9B%A2> score: 0.1840358674

国家会议中心_百度百科 url: <https://baike.baidu.com/item/%E5%9B%BD%5E%5A%8E%4E%BC%9A%5E%8A%5E%4B%8A%AD%5E%BF%83> scor

家乐福_百度百科 url: https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#F1_2 score: 0.183576643467 contents: I

家乐福_百度百科 url: https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#F8_1 score: 0.183576643467 contents: I

家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#ref> [13]_18119 score: 0.183576643467

家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#ref> [7]_18119 score: 0.183576643467 (

家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#F1> score: 0.183576643467 contents: Nor

家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#ref> [2]_18119 score: 0.183576643467 (

家乐福_百度百科 url: https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#F1_3 score: 0.183576643467 contents: I

家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#F7> score: 0.183576643467 contents: Nor

家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#ref> [11]_18119 score: 0.183576643467

家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#ref> [8]_18119 score: 0.183576643467 (

家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F> score: 0.183576643467 contents: None

家乐福_百度百科 url: https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#F1_4 score: 0.183576643467 contents: I

家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F#ref> [5]_18119 score: 0.183576643467 (

仔细看这些网页的url是不同的：末尾的页码信息。像百度百科这种网站，随着用户往下拉滚动条，触发js动作，于是动态加载下面的新内容，而在这一过程中，url的尾部页码信息就会改变了。爬虫时我们判断某网页是否已爬取是用url作为Primary Key的，所以这里没能做到重复筛查。所以在搜索时，要以title为Primary Key做一次重复筛查。

title: 家乐福_百度百科 url: <https://baike.baidu.com/item/%E5%AE%86%4B%90%9E%7%A6%8F%E9%9B%86%5E%9B%A2> scor

title: 北京91装修-新浪家居网-新浪家居网 url: <http://91.jiaju.sina.com.cn/> score: 0.253758341074 contents: Nor

title: 腾讯携手永辉入股家乐福中国_百科TA说 url: <https://baike.baidu.com/tashuo/browse/content?id=38aea596cec1>

title: 最新最全的“福克斯”资讯都在这里，不容错过 url: <http://auto.sina.com.cn/autonews/zt/2018-10-17/detail-ifx>

title:

家乐福_历史版本_百度百科 url: <https://baike.baidu.com/historylist/%E5%AE%86%4B%90%9E%7%A6%8F/354372> score:

title: 福建师范大学_百度百科 url: <https://baike.baidu.com/item/%E7%8A%8D%5A%8E%4E%BC%9A%5E%BF%88%5E%9B%A2>

title: 福利经济理论_百度百科 url: <https://baike.baidu.com/item/%E7%8A%8D%5A%8E%4E%BC%9A%5E%BF%88%5E%9B%A2>

title: 福利经济学（经济学分支学科）_百度百科 url: <https://baike.baidu.com/item/%E7%8A%8D%5A%8E%4E%BC%9A%5E%BF%88%5E%9B%A2>

title: 盘点娱乐圈的模范夫妻_新浪视频话题_新浪网 url: <http://video.sina.com.cn/topic/100003318.html> score: 0.17

title: 福建幼儿师范高等专科学校_百度百科 url: <https://baike.baidu.com/item/%E7%8A%8D%5A%8E%4E%BC%9A%5E%BF%88%5E%9B%A2>

title: 上海演出搜索-秀动网演出列表页 url: <http://www.showstart.com/event/list> score: 0.168891429901 cont

title: 西奥多·罗斯福_百度百科 url: <https://baike.baidu.com/item/%E8%A5%BF%5A%5E%5A%49%A%2C%7%BD%97%5E>

12 total matching documents.

筛查后的搜索结果

②爬到pdf、下载文件

在建立索引时发现对某些文档索引耗时非常长，分析后发现是爬到了pdf文件、下载命令等。我直接把这些文件当成网页，储存为了html格式的文件。而写入文件其实都是写入二进制码，也就是把整个文件的二进制码写入了html文件，该文件占用了很大的磁盘，并且索引时需要对大量无意义的文本信息进行搜索，耗时巨大。因此在爬虫时应该对这些情况进行初筛；在建立索引的时候同样应该进行二次筛选。



pdf以‘utf-8’被写入文件，大量无意义乱码

③关于中文编码

在Python文件头声明了# -*- coding: utf-8 -*-，并且BeautifulSoup是默认使用utf-8编码的，也就是

对网页decode之后再统一转换成utf-8，按道理是不应该再出现编码问题的。然而还是出现了少量的编码问题（约15000个网页中出现10次），如下图所示。

```
adding http://baby.sina.com.cn/healhmmjkhhyq2018-10-17doc-ifxeuwzs  
孕期看牙 局麻不是禁忌|孕妇|胎儿|麻药_新浪育儿_新浪网  
1134  
adding http://finance.sina.com.cn/fund/company80064562.shtml  
巴西世界杯赛程表_巴西世界杯赛程表_新浪竞技风暴_新浪网  
1135  
adding https://sports.sina.com.cn/others/volleyball2018-10-15doc-ihm1  
女排：中国男排正处艰难时刻 输球也不能输掉信心_排球_新浪竞技风暴_新浪网  
1136  
adding http://baike.baidu.com/item/F8AFRAFRR49NF5R094F5A596vhaikenz
```

调用soup.original_encoding查看网页的原编码，发现使用‘gb2312’等编码的网页都被正确地转成了‘utf-8’，出现问题的都是采用了‘Windows-1252’编码的网页。使用bs中的.prettify()语句无果，我又手动对文本先解码再转码，.decode('windows-1252').encode('utf-8')，却报错‘无法decode’，或许是该字符串在前序过程中已经不是按原编码储存了。考虑到出现频率较小，最终暂未解决。

3.组合搜索

3.1.原理简述

组合搜索在生活中十分常用。比如搜索某首歌曲，想要直接拿到文件，可以使用“周杰伦 file:mp3”搜索；比如搜索某个法律条款，为了保证权威性和官方性，可以使用“税率 site:gov”搜索。首先我们要将对网页的site建立索引，然后在搜索时要对用户的输入进行预处理。

3.2.site用法

刚开始我直接将文档的url储存为site并索引，搜索时返回url中含有site字段的文档即可。然而实际使用时，总是会混入并不在当前site的内容，这种字符串包含匹配的方式并不准确。

于是我使用百度搜索的site:方法，发现了site:使用的一些规则：

①不可以带有'/'，'http://'等内容。

②基于①，site只是url在第一个'/'以前的内容，也就是网址的站点。可以使用`urlparse.urlparse(url).netloc`语句获得。

很抱歉，没有找到与“壁纸 site:image.baidu.com/channel/wallpaper”相关的网页。

site内容中带有'/'，不合法

③site右连续性。比如www.image.baidu.com，其site

有'www.image.baidu.com'，'image.baidu.com'，'baidu.com'，'com'。因此，www.baidu.com和baidu.com都是合法的site，但是使用效果并不同。很明显后者涵盖的范围更大。运用这一方法，我们也可以使用site来控制后缀名来限定网页搜索的范围。如下图所示：

④site不可以叠加（一次搜索只能用一个site）。

site命令不可以叠加

⑤site命令前后都可以输入内容，以空格为间隔。

The image shows two side-by-side search results pages from Baidu. The left page is for '中国 上海 site:gov' and the right is for '中国 site:gov 上海'. Both pages show results for 'Buyusa.gov' and 'ICE.gov'. The Chinese version includes links to 'ShareAmerica.gov' and 'ClinicalTrials.gov'. The English version includes links to 'USAID.gov' and 'USA.gov'. The results are identical in both versions.

3.2.结果展示

对照着上方所述的我发现的规律，对每个文档我加入了site域名，以索引且保存的方式。对搜索命令同样进行了预处理，在代码中有注释，此处不再赘述。

Top ranking terms. (Right-click for more options)			
Rank	Freq	Field	Text
1	9182	site	cn
2	9050	site	com.cn
3	8995	site	sina.com.cn
4	5646	site	com
5	4802	site	baidu.com
6	3643	site	baike.baidu.com
7	1273	site	finance.sina.com.cn
8	1085	site	auto.sina.com.cn
9	1066	site	sports.sina.com.cn
10	1050	site	tech.sina.com.cn
11	839	site	ent.sina.com.cn
12	810	site	blog.sina.com.cn
13	536	site	news.sina.com.cn
14	464	site	wenku.baidu.com
15	314	site	video.sina.com.cn
16	273	site	slide.ent.sina.com.cn
17	247	site	v.baidu.com
18	236	site	db.auto.sina.com.cn
19	228	site	sina.com
20	227	site	med.sina.com
21	221	site	fashion.sina.com.cn
22	188	site	weibo.com
23	188	site	astro.sina.com.cn
24	170	site	weather.sina.com.cn
25	169	site	fo.sina.com.cn
26	146	site	stock.finance.sina.com.cn
27	143	site	vip.stock.finance.sina.com.cn
28	139	site	k.sina.com.cn
29	137	site	mil.news.sina.com.cn
30	132	site	communist.sina.com.cn
			['aonovi.sina.com.cn', 'sina.com.cn', 'com.cn', 'cn']
			s.weibo.com
			['s.weibo.com', 'weibo.com', 'com']
			weibo.com
			['weibo.com', 'com']
			my.sina.com.cn
			['my.sina.com.cn', 'sina.com.cn', 'com.cn', 'cn']
			weibo.com
			['weibo.com', 'com']
			gongyi.sina.com.cn
			['gongyi.sina.com.cn', 'sina.com.cn', 'com.cn', 'cn']
			gongyi.sina.com.cn
			['gongyi.sina.com.cn', 'sina.com.cn', 'com.cn', 'cn']
			gongyi.sina.com.cn
			['gongyi.sina.com.cn', 'sina.com.cn', 'com.cn', 'cn']
			gongyi.sina.com.cn
			['gongyi.sina.com.cn', 'sina.com.cn', 'com.cn', 'cn']
			gongyi.sina.com.cn
			['gongyi.sina.com.cn', 'sina.com.cn', 'com.cn', 'cn']
			s.weibo.com
			['s.weibo.com', 'weibo.com', 'com']
			gongyi.sina.com.cn
			['gongyi.sina.com.cn', 'sina.com.cn', 'com.cn', 'cn']
			gongyi.sina.com.cn
			['gongyi.sina.com.cn', 'sina.com.cn', 'com.cn', 'cn']
			gongyi.sina.com.cn

建立的site索引

Searching for: 中国
title:
开始大刀阔斧会师新疆——当前中国改革发展述评之四——新华网
url: http://www.xinhuanet.com/politics/2018-10/11/c_1123543563.htm score: 0.281845659018 contents: None
title: 中国的和平发展 百度百科 url: <https://baike.baidu.com/item/%E4%BB%A3/%E5%9B%9B/%E7%A8%8B%C5%8B%83%EF%91%95/B19195> score: 0.255676686764 content:
title: 金子数据头条·年度·财经新闻 新浪财经 url: https://cs.sina.com.cn/k/article/author_article_id/2622742392.html score: 0.255296945572 contents: None
title: 苏联局势动荡全部新闻 新浪军事 新浪网 url: https://roll.mil.news.sina.com.cn/z_xohifxdy/all/index.shtml score: 0.25416713953 contents: None
title: 苏联解体 [中央美术学院附属画廊] 百度百科 url: <https://baike.baidu.com/item/%E5%9B%9B/E6%90%80/E5%BF%93/5792596> score: 0.249880379641 contents: None
title: 微信注册 [微博] url: <https://weibo.com/signup/signup.php> score: 0.246417045593 contents: None
title: 美国女伶统治女排后将迎死神! 美国奥运迷魂团到了李盈莹 女排队 [中国女排] 豆瓣网 新浪网 url: https://k.sina.com.cn/article/2419389333_9033b1501000ef5ef.html?from=weibo score: 0.244617045593 contents: None
title: 德国学者:美国想让中国女排没落 针对普吉白邦瑞 新浪军事 新浪网 url: <https://mil.news.sina.com.cn/2018-10-17/doc-ifxewns5191057.shtml> score: 0.244533807039 contents: None
title: BWC中文词条条主页 [财经类] 新浪财经 新浪网 url: https://c.ji.sina.com.cn/k/article/author_article_id/19865347354 score: 0.244533807039 contents: None
title: 中央美术家协会 url: <http://www.caanet.org.cn/> score: 0.243583176187 contents: None
title: 中华全国归国华侨联合会 url: <http://baike.baidu.com/view/22114.htm> score: 0.225186407566 contents: None
title: 墓碑“耀”发 中国古支激光雕刻墓碑网“耀” url: <http://baijiahao.baidu.com/s?id=161394671754463055> score: 0.224507927895 contents: None
title: 中国书画家协会 url: <http://www.cagov.org.cn/> score: 0.221647381783 contents: None
title: 中国世界四大文明古国之一 url: <http://baike.baidu.com/view/391361.htm> score: 0.221564918756 contents: None
title: 林森生 百度百科 url: <https://baike.baidu.com/item/%E6%96%95/%E5%AE%BD/%E7%94%9F> score: 0.221526065528 contents: None
title: [网上网下·辉煌·十九大] 光明专题 url: <http://topics.gmw.cn/node/114815.htm> score: 0.221086236357 contents: None
title: 国际投资论坛专题论坛:中医药国际化创新发展之路 新浪网 url: http://finance.sina.com.cn/zt/d_zqgtzmyth/ score: 0.219460397959 contents: None
title: 医疗器械 百度百科 url: <https://baike.baidu.com/item/science/medical/device> score: 0.219312891364 contents: None
title: ammin001 新浪博客 url: <http://blog.sina.com.cn/u/1253386310> score: 0.218855112791 contents: None
title: 冷军萧鼎:中国不能不懂美国 恶习无半点便知.手机新浪网 url: http://news.sina.com/global/2018-10-10-detail-ihkvhrpt4745984_d.html score: 0.218623757362 content:
title: 爱心大使·中国汽车领军人物汽车社区·汽车资讯·汽车论坛中企网 url: <http://www.xcar.com.cn/> score: 0.218551829457 contents: None
title: 中国中医药网 url: <http://www.cntcm.com.cn> score: 0.217631220188 contents: None
title: 新丝绸之路——一带一路国际信息服务平台.丝绸之路经济带与21世纪海上丝绸之路权威网站 url: <http://silkroad.news.cn/> score: 0.212658584118 contents: None
title: 王平庆:海上丝绸之路的内涵、形成与演变 王平庆 新浪博客 url: http://blog.sina.com.cn/s/blog_487ad3101820tpp.html score: 0.2122026442089 contents: None
title: 武连生 中国留学网 学院办创新能力提升榜 榜单 新浪博客 url: http://blog.sina.com.cn/s/blog_482c00e0102xpwv.html score: 0.211032770205 contents: None
title: 用什么来化解影射式的“焦虑”.中国经济网——国家经济门户 url: http://www.ce.cn/cw/zx/gnzs/gndw/2018/10/11/20181011_30488933.shtml score: 0.209105268121 contents: None
title: [九大·理论视点] url: <http://www.gstheory.cn/zt2017/xcgdd19djs/index.htm> score: 0.208617851138 contents: None
title: 中国石油石化集团公司.百百科 url: <https://baike.baidu.com/item/%E9%8B%8A%E5%9B%BD/%E5%8B%8B#sef=rcr> score: 0.207190662622 contents: None
title: 孙中山《日光盟约》与世界第2季是一回事吗.相关新闻 url: http://blog.sina.com.cn/s/blog_485dc670102ye7b.html score: 0.204817205371 contents: None
title: 武连生 中国留学网 学院办学系评价榜·武林榜 新浪博客 url: http://blog.sina.com.cn/s/blog_482c00e0102xpfn.html score: 0.204651635767 contents: None
title: 大学生排行榜指南用指南 武连生 新浪博客 url: http://blog.sina.com.cn/s/blog_482c00e0102xpjw.html?item=title&id=711 score: 0.203772589564 contents: None
title: PUPHC中国植物数据库——最具代表性的植物类图片网 url: <http://www.plantphoto.net> score: 0.203237310052 contents: None
title: 跳跳乐.新浪微博 url: <http://blog.sina.com.cn/u/1724710545> score: 0.202150518488 contents: None
title: 赵晶晶 百百科 url: <https://baike.baidu.com/item/%E8%8C%83/%E5%8D%80/%E5%8B%95> score: 0.2019803298445 contents: None
title: 中国寻求新刑法法典 大幅增加处罚力度 新闻观察网 url: http://med.msn.com/article_detail_2_2_3417.html#p_l-listbox score: 0.198093298445 contents: None
title: 中国男篮 [中国男子篮球职业联赛] 百百科 url: <https://baike.baidu.com/item/%E4%BA%A4/%E5%8D%80/%E5%8B%95> score: 0.1985893956 contents: None
title: 中国国家博物馆 [中国国家博物馆] 百百科 url: <https://baike.baidu.com/item/%E4%BA%A4/%E5%8D%80/%E5%8B%95/%E7%8E%89/%E9%9A%8E/%E6%9C%89> score: 0.198032770461 contents: None

直接搜索“中国”

Searching for: 中国 site:baike.baidu.com

title: 中国的和平发展_百度百科 url: <https://baike.baidu.com/item/%E4%B8%AD%E5%9B%BD%E7%9A%84%E5%92%8C%E5%B9%B3%E5%8F%91%E5%81%95> score: 1.11344718933 contents: None

title: 苏新平(中央美术学院副院长) _百度百科 url: <https://baike.baidu.com/view/22114.html> score: 1.1134052277 contents: None

title: 中华民国_百度百科 url: <http://baike.baidu.com/view/1893161.htm> score: 1.1136293411 contents: None

title: 林容生_百度百科 url: <https://baike.baidu.com/item/%E6%9E%97%E5%AE%89%E7%94%9F> score: 1.1134052277 contents: None

title: 中国石油化工集团公司_百度百科 url: <https://baike.baidu.com/item/%E4%BB%AD%E5%9B%BD%7F%9F%83%5%8C%96/153933> score: 1.10383343697

title: 裴建勋_百度百科 url: <https://baike.baidu.com/item/%E6%AF%95%E5%8B%BA%E5%8B%8B%8B%8B?sef=cr> score: 1.10309016705 contents: None

title: 中国绿卡_700万人梦想_60万人申请,但每人只能领一个诺贝尔奖_百科TA说 url: <https://baike.baidu.com/tashuo/browse/content?id=63ebc0fd574>

title: 吴为山_百度百科 url: <https://baike.baidu.com/item/%E5%99%A8%E5%8A%9B%F5%86%F5%80%9A%F7%80%9A%F9%A6%86/567902>

title: 中国足球协会超级联赛_(中国大陆职业足球联赛)_百度百科 url: <http://baike.baidu.com/search/word?word=%E4%B8%AD%E0%80%85> score: 1.097984

title: 毛泽东 (无产阶级革命家)_百度百科 url: <https://baike.baidu.com/item/%E6%AF%95%E6%83%BD%4E%8B%9C> score: 1.0977069092 contents: None

title: 华中师范大学_百度百科 url: <https://baike.baidu.com/item/%E5%8D%8E%E4%BB%AD%E5%8B%88%E8%C8%83%5%A4%A7%5%AD%86> score: 1.09676110

title: 姚永 (油画家)_百度百科 url: <https://baike.baidu.com/item/%E5%A7%9A%E6%80%80/1964643> score: 1.0966540575 contents: None

title: 国际经济学 (指研究国民经济活动和国际经济关系)_百度百科 url: <https://baike.baidu.com/item/%E5%9B%BD%F9%98%85%7F%BB%8F%6B%8E%5%AD%86>

title: 为幸福而奋斗!纪念改革开放40周年专题展在京举办_百科文章_百度百科 url: <https://baike.baidu.com/article/b66f18449dcaf980/f007e.htm>

鲁迅美术学院与索尔福德大学合作办学签约仪式举行_百科文章_百度百科 url: <http://baike.baidu.com/article/2767hec45dc7bac70673670.htm>

title: 亚洲二十世纪及当代艺术香港秋拍_百科文章_百度百科 url: <https://baike.baidu.com/article/936cd57105cb50d060412a7e.htm> score: 1.09437572

title: 中国国家女子排球队_百度百科 url: <https://baike.baidu.com/item/%E4%BB%AD%E5%9B%BD%5CAF%86%5A5%AD%90%E6%8E%92%E7%>

title: 宪法(法律名)_百度百科 url: <http://baike.baidu.com/view/3575.htm?fbid=b5060816> score: 1.09167063236 contents: None

title: 改革开放_百度百科 url: <https://baike.baidu.com/item/%E6%98%F9%9D%A9%5E%8C%80%80%94%BC> score: 1.09137153625 contents: None

title: 《北京的艺术毕业季》活动暨“全国高校联盟”开幕_百科文章_百度百科 url: <http://baike.baidu.com/article/d276f17ce5c75e/6870.htm> score: 1.09050011635 cc

title: 佳士得秋拍 汇集东西艺术_百科文章_百度百科 url: <https://baike.baidu.com/article/edc68a069a61211fb6a17d.htm> score: 1.09050011635 cc

title: OFII_百度百科 url: <http://baike.baidu.com/view/24777.htm> score: 1.09033453465 contents: None

title: 国家会议中心_百度百科 url: <https://baike.baidu.com/item/%E5%9B%BD%5AE%86%4E%BC%9A%8E%AE%8E%4B%8F%83> score: 1.08993136

title: 65幅嫡本精美唐卡作品亮相国博_百科文章_百度百科 url: <https://baike.baidu.com/article/15c17d1422ed7f0ca8410f7e.htm> score: 1.089861512

宗其善百年诞辰纪念展: 画家怎么把画笔当成武器_百科文章_百度百科 url: <http://baike.baidu.com/article/71260de10cb244fc90d31370.htm> score: 1.089861512

title: 万方数据知识服务平台_百度百科 url: <https://baike.baidu.com/item/%E4%BB%AD%5E%8B%87%F6%6A%89%F6%95%80%F6%8B%AF%7%9F%A5%E%8B%86%E%9C%8D%E>

轴心国_百度百科 url: <http://baike.baidu.com/view/32267.htm> score: 1.08792495728 contents: None

title: 百度科学百科 url: <https://baike.baidu.com/science> score: 1.0878977756 contents: None

赛珍珠_百度百科 url: <https://baike.baidu.com/item/%E8%8B%9B%8E%7F%8P%8D%8E%7F%8P%80> score: 1.08757913113 contents: None

title: “比格尔与当代艺术”学术论坛_6/24 (周六) | 青年思想者驻馆项目_百科文章_百度百科 url: <http://baike.baidu.com/article/ab3b24ecd3d006e96a4d9>

title: 2008年北京奥运会_百度百科 url: <http://baike.baidu.com/view/16667.htm> score: 1.0869412301 contents: None

title: 中国知网_百度百科 url: <https://baike.baidu.com/item/%E4%BB%AD%5E%9B%BD%F7%AF%5A%5E%BD%91/1316830> score: 1.08506655693 contents: None

title: 韩洪伟 (画家)_百度百科 url: <https://baike.baidu.com/item/%E4%9F%9D%E5%84%84%F4%BC%9F/2205654#viewPageContent> score: 1.08490288

title: 中央银行_百度百科 url: <http://baike.baidu.com/view/79768.htm?hold=redirect> score: 1.0844495295 contents: None

title: 国瓷·红叶系列产品荣获2018巴拿马金奖_百科文章_百度百科 url: <https://baike.baidu.com/article/0cbc6fa2bccfc938f20e047e.htm> score: 1.084

在“baike.baidu.com”搜索“中国”

Searching for: 中国 site:sina.com.cn

title: 2018俄罗斯世界杯_新浪体育_新浪网 url: <http://2018.sina.com.cn/> score: 0.700619339943 contents: None

title: 赛车_新浪体育_新浪网 url: <http://f1.sina.com.cn/> score: 0.696150779724 contents: None

title: 金十数据头条_财经头条_新浪财经 url: <https://cj.sina.com.cn/k/article/author/article/2622472937> score: 0.689533948898 contents: None

title: 美国队输给中国女排后将迎生死战! 美国球迷重点提到了李盈莹_美国队_李盈莹_新浪网 url: https://k.sina.com.cn/article_2419309333_9033bb150010

title: BWC中文网头条_财经头条_新浪财经 url: <https://cj.sina.com.cn/k/article/author/article/1986534745> score: 0.68148291111 contents: None

title: 国际投资论坛专题论坛_中医药国际化创新发展之路_新浪网 url: <http://finance.sina.com.cn/zt/d/gqjzmyzh/> score: 0.662727594376 contents: None

title: annin_0001_新浪博客 url: <http://blog.sina.com.cn/u/1253386310> score: 0.662274837494 contents: None

title: 王庆丰: 海上丝绸之路的内涵, 形成与演变_王庆丰_新浪博客 url: http://blog.sina.com.cn/s/blog_a0d1102v0htp.htm score: 0.65731215477 content

title: 武书连中国独立学院大学创新能力排行榜_武书连_新浪博客 url: http://blog.sina.com.cn/s/blog_4b2cbb00e102xpz/w_.html score: 0.65642362833 content

title: 孙中英(中盟约)与孙世凯21条是一回事_程万里_新浪博客 url: http://blog.sina.com.cn/s/blog_4b2cbb00e102ve7b_.html score: 0.651774287224 content

title: 武书连中国独立学院大学升学率排行榜_武书连_新浪博客 url: http://blog.sina.com.cn/s/blog_4b2cbb00e102xp0f_.html score: 0.651650428772 counter

title: 大学排行榜实用指南: 分层次报考_武书连_新浪博客 url: http://blog.sina.com.cn/s/blog_4b2cb00e102x1ui.html?tj=edu&tj1 score: 0.650992870331 (

title: 烟树浩_新浪博客 url: <http://blog.sina.com.cn/u/1724710054> score: 0.6497797966 contents: None

title: 胡曼杰_新浪博客 url: <http://blog.sina.com.cn/u/1215347692> score: 0.647087037563 contents: None

title: 战后降落中国的日本飞机有多少?_郭林_新浪博客 url: http://blog.sina.com.cn/s/blog_470fe5b0102v54u.html score: 0.643809020519 contents: None

title: 5天4连胜。女排10年宿敌军窝囊出局, 泽谭国双杀他们点太背! 盖世冠军_宿敌_美国队_新浪网 url: https://k.sina.com.cn/article_6348157290_17a61316a0

title: 出国频遭|出国留学 留学签证 留学费用 世界大学排名 移民_新浪教育_新浪网 url: <http://edu.sina.com.cn/a/> score: 0.634827971458 contents: None

title: 《经经》隐藏了中国人的什么秘密?_史财经_新浪网 url: <http://www.sina.com.cn/zhihixing/video/2015-11-24/doc-ifxkwuwu3582827.shtml> score: 0.63235776165 contents: None

title: 古迹中国_新浪博客 url: <http://blog.sina.com.cn/u/1235158341> score: 0.63235776165 contents: None

title: 世纪大讲堂_新浪博客 url: <http://blog.sina.com.cn/u/1223515835> score: 0.630647301674 contents: None

title: 程万里_新浪博客 url: <http://blog.sina.com.cn/u/1214106727> score: 0.630647301674 contents: None

title: 国企“竞争中性”治愈金融业顽疾_冉学东_新浪财经 url: http://blog.sina.com.cn/s/blog_617c39a40102vtb_.html score: 0.630161523819 contents: None

title: 梁建章_新浪博客 url: <http://blog.sina.com.cn/u/1925476280> score: 0.630088329315 contents: None

title: 刘亦菲凭什么不能演花木兰? 但真人版《花木兰》不管谁演都差成功_郎罗大电影_新浪博客 url: http://blog.sina.com.cn/s/blog_1515cbfd80102x5ea.html?tj=1 score: 0.62748431199 contents: None

title: 坐山观虎斗! 女排有望前进四强, 猜蜜赛程成冲冠另一帮手_美国女排_女排_中国女排_新浪网 url: https://k.sina.com.cn/article_1315856454_4e6e60460010

title: 许晖_新浪博客 url: <http://blog.sina.com.cn/u/1867153750> score: 0.621709764804 contents: None

title: 首届丝路国际产能合作论坛_新浪财经 url: <http://finance.sina.com.cn/zt/d/sjgjcnhz1/> score: 0.626205503941 contents: None

title: 吕培坚_新浪博客 url: http://blog.sina.com.cn/s/blog_f0425e0301ia.html?tj=1 score: 0.624535083771 contents: None

title: 泰国旅游_泰国旅游攻略_景点_视频_微博_图片_自由行线路推荐_新浪旅游 url: <http://travel.sina.com.cn/taiquo-lvyou/> score: 0.622242808342 contents: None

title: 意大利球迷炮轰世锦赛裁判不公: 凭啥是咱中国女排最后一天打客场? [荷兰] 晚场_意大利_新浪网 url: https://k.sina.com.cn/article_5457712446_1454575e0010

title: 世界女排格局大变! 老牌劲队美巴俄集体出局, 欧洲三强挑大梁! [美国]女排队_强队_女排_中国女排_新浪网 url: https://k.sina.com.cn/article_6348157290_17a61316a0010

title: 肖森新市_新浪博客 url: <http://blog.sina.com.cn/u/1192082717> score: 0.621177925493 contents: None

title: 钮文新_新浪博客 url: <http://blog.sina.com.cn/zt/> score: 0.620662331581 contents: None

title: 体育专题汇总_新浪竞技风暴_新浪网 url: <http://sports.sina.com.cn/zt/> score: 0.620107233524 contents: None

title: 张化桥_新浪博客 url: <http://blog.sina.com.cn/u/135520384> score: 0.619835734367 contents: None

title: 德国学者: 美国连伊拉克都没法重建了, 怎么重建中国_环球时报_特朗普_白邦瑞_新浪军事_新浪网 url: <https://mil.news.sina.com.cn/2018-10-17/doc-ifxeuwws>

title: 知否: 周末举行的中国经济50人论坛全貌就在那里了_新浪网 url: <http://finance.sina.com.cn/zt/d/50ren50th/> score: 0.618785262108 contents: None

在“sina.com.cn”搜索“中国”

Searching for: 中国 上海 site:edu.cn

title: 中央美术学院 url: <http://www.cafa.edu.cn/> score: 4.61260128021 contents: None

title: 清华大学美术学院 url: <http://www.tsinghua.edu.cn/publish/ad/index.html> score: 4.61120843887 contents: None

2 total matching documents.

右连续性, 并且搜索命令contents可以分开

4.图片爬虫

4.1.任务简述

最终我们要做到的是: 输入文本, 可输出相关的图片地址, 图片所在网页的网址, 图片所在网页的标题。整个任务可以分解成几个部分:

①爬取网页

②定位网页中的图片

③分析结构信息, 对图片加上文本信息

④以图片为Primary Key，处理文本信息并建立索引

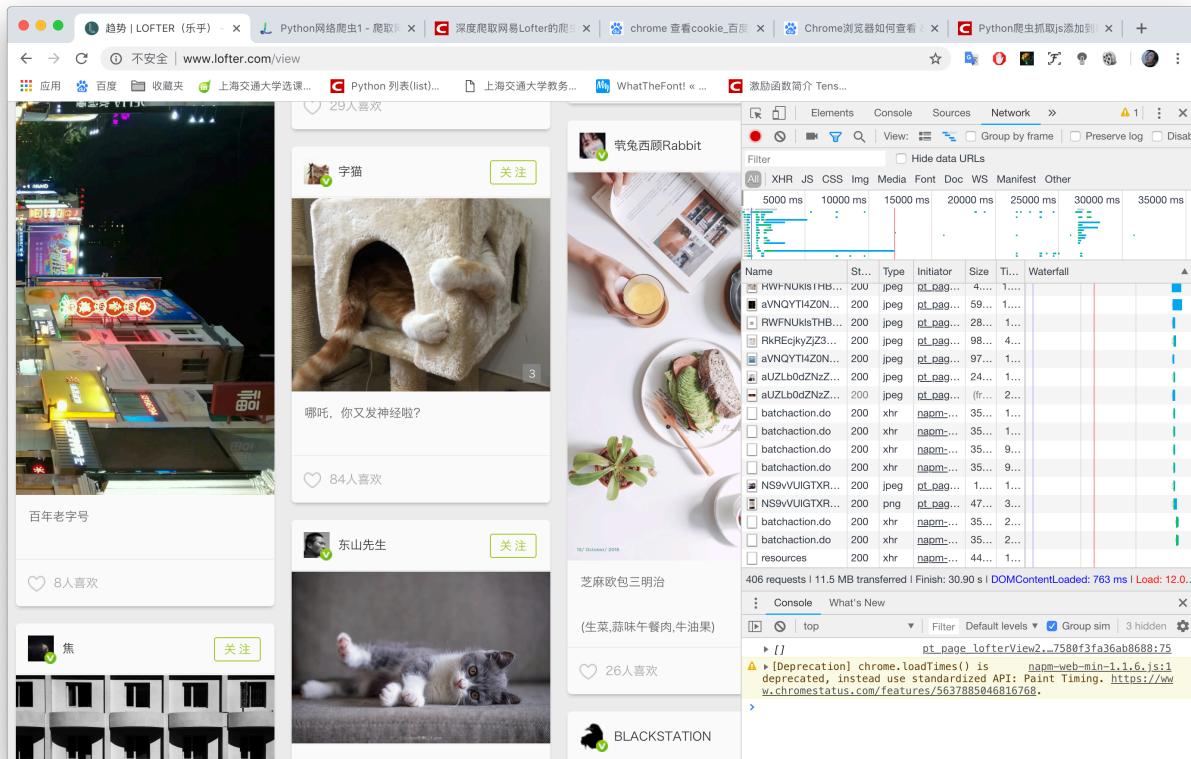
①③和④在之前的实验中都有完善的解决方案，稍做修改就可以使用。这次实验我遇到的最大难点来自任务②。爬取页面中的图片绝不仅仅是搜索'img'标签，而是需要结合动态网页的爬取、抓包等。这为我打开了新世界的大门。

4.2.实例：爬取动态网页

要爬取丰富的图片内容，我首先就想到了lofter.com这一被称作“中国ins”的图片分享网站，上面有大量的来自摄影师和设计师的高品质图片。携带着含有我登录信息的cookie用常规的

`urllib.request.urlopen()`打开“lofter.com/view”这一浏览主页，返回该网页内容，并在内容中搜索标签为'img'的节点，却发现一个网页中只有30来个img，并且都是logo、图片缩略图等品质较差的图片，显然达不到爬取目的：高品质图片。而对比人工查看的网页，一张图本来显示的缩略图，要点击才能加载出高清大图；并且网页刚开始只显示少量图片，随着用户往下滚动页面，更多图片才会被加载出来。而在这整个过程中，url并没有改变。

意识到：这是一个动态网页。用户进行了特定的动作后，会触发相应的js命令，网页再传回来储存着更丰富信息的包，我们要抓取这些包里的内容。



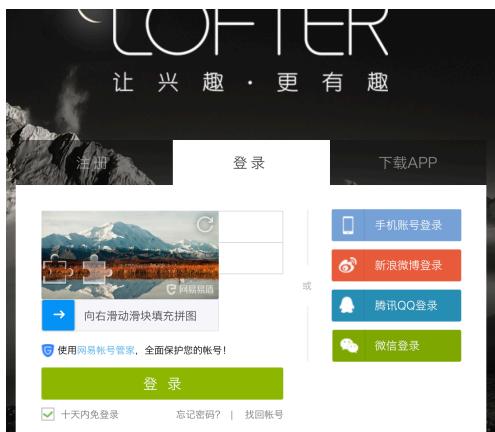
使用XHR对象，和服务器进行交互

那么如何处理动态网页呢？使用之前的`urllib.request.urlopen()`方法，只能传回打开url那一刻的信息，之后的信息就无能为力了。前文在讨论爬取url重复时，提到了百度百科的动态加载功能，用户下滑，发送js命令，url被改变，刷新新页面。那么要获得最终页面，只需要想办法解析出最终url，打开即可。然而，lofter用的不是这种方法，而是不改变url、不刷新网页，随着用户操作实时传回包，也就是实时改变网页信息，当前网页=原网页+传回包。

经过网上自学异步爬虫相关教程，我首先使用了 PyQt5 库。PyQt5 库常用来编写 Python 脚本的应用界面，在这里应用到异步爬虫，可以模拟人工访问网页，爬到动态网页加载完 js 的内容。可是运行之后我发现，爬到的内容仍很少，这是因为 lofter 的动态 js 并不是随着时间可以自动加载完的，而是需要人工点击、滑动等操作。如何模拟人工的点击、滑动、拖拽等动作呢？

我了解到 Selenium 库。Selenium 是一个浏览器自动化测试框架，当然也可以在这里模拟浏览器。我首先为其安装了 Chrome 的驱动，以便于观察实际操作；事实上，最终爬虫使用的是 phantomJS，一个无头浏览器，更加快速。

用 Selenium 模拟访问，首先遇到的困难是登录。找到页面中用户名、密码对应的文本框，填入我的信息，这都很简单。然而不是直接能用 POST 发送的，还需要滑动验证码。



纵然 Selenium 可以模拟拖拽，但如何识别滑动到了正确的位置呢？对于计算机就需要用到图像识别了，这次实验中我不想更加复杂，于是先在 Selenium 中人工登录，`.get_cookie()` 获得登录后的 cookie 信息，储存到本地。之后每次 Browser 打开 lofter 站点下的网站时，就先打开本地的 cookie 文件，逐个`.add_cookie()`，再刷新网页，显示的就是已登录界面了。

在登陆后的浏览页面，首先全是用户动态的小图。要查看大图就得先点击小图，要点击小图就得先定位到小图，而定位小图则需要先对当前页面信息发掘，找到能精确定位到小图的方式。在这个过程中，xpath 语法非常好用。xpath 同时可以定位当前节点的父节点、祖先节点、兄妹节点等，对提取图片周围信息有用。然后`.onclick()` 操作，点开小图，进入画廊浏览模式，实际是传回了 xhr 包，其中包含着高清大图的地址、图片信息描述等，将这些信息传回本地，就实现了爬虫。此时再定位到精确对象（如大图、标签、描述）进行`.get_attribute()`，就可以返回图片信息了。接下来，定位到返回叉或者‘下一张’按钮，点击‘叉’关闭窗口回到主页，或者点击‘下一张’来看下一张大图。

在报告附件中有操作录屏的 gif，可以看到 Selenium 用 Chrome 驱动自动完成了登录、点开、下一张、储存等操作。但是仍然有两个无法解决的问题。

其一非常慢。由于是模拟人工操作，操作时间需要大于 js 库中设定的人体动作阈值才有效，并且等待网页传回包，动态加载，是很耗时间的。在测试代码中我使用了`time.sleep()` 方法，实际上可以用`expected_conditions` 配合`WebDriverWait` 等待加载完毕。但仍然很慢。

其二是网站的反爬虫机制。从 gif 中可以看到，爬了 4 张图片后就报错‘无法定位到元素’而导致窗口关闭而停止。网页结构都一样，同样的 xpath 语句，怎么会无法定位元素呢？说明网页没有返回这一元

素或者在某个时刻之后才会返回。事实上，我第一次运行该程序时连续爬了20多张后报错，之后能连续爬的数量越来越少。这说明我的ip或者用户已经被反爬虫机制所监测到并进行了一定程度的封禁。也可以理解，这种用户分享型网站，为了保护自己的数据和用户版权，反爬虫机制必然鲁棒。在网上查这种莫名报错，也有反映是由于反爬虫。比如人类在点击‘下一张’浏览时，不会每次点击的时间间隔、位置严格相同，因此还需要学习反反爬虫的策略，比如对间隔时间、点击位置进行设置随机的微调offset等。但是这次，爬取动态网页和抓包的基本方法框架已经设置好了，暂时收手，下次再研究反爬虫。

4.3.爬取静态网页

前面的尝试爬取动态网页使我精力耗尽，因此再使用常规操作爬取静态网页时，我决定使用较简单的方法。要爬取图片，重点在于分析出图片周围的信息以描述这张图片。以我现在的能力，最好是爬取同一结构的网页，这样提取图片信息较为准确。因此我选择爬取京东的商品页，该页面有两个好处，其一是大体结构固定，可以容易地定位到图片，其二是title内容足够丰富并且足够准确描述图片。接下来分析我们的目的，我们并不是要定向找出和某个内容有关的图片，而是要先尽可能多而广地爬取图片，再对之进行索引。这样，如果再使用之前的获取页面所有超链接的方式爬取，由于京东的商品推荐，爬取内容会大量同质化，不符合我们的需求。再分析京东商品页的url构造，都是“[https://item.jd.com/”+itemNo+.html”，这样，一种极其简单但是十分有效的方式应运而生——随机生成大量的不重复的七位数作为itemNo放入url中，爬取即可。或许会遇到没有商品的情况，但是无妨。在大量爬取（10000个商品，耗时3分钟）的情况下，最终爬取到的商品覆盖的范围足够广。](https://item.jd.com/)

接下来是对图片的信息进行索引和搜索。过程和前文对文字网页索引和搜索的过程并无二致，此处不再赘述。展示如下：

```
Searching for: 陶瓷
图片: https://img12.360buyimg.com/n1/jfs/t412/246/58017148/34390/ac2f76c5/5407e0feNad2d2766.jpg
名称: 【美瓷陶瓷菜刀】美瓷(MYCERA)陶瓷菜刀 (黑色) AHG6.5B
链接: https://item.jd.com/1208647.html
相关度: 3.32669758797

图片: https://img12.360buyimg.com/n1/jfs/t412/246/58017148/34390/ac2f76c5/5407e0feNad2d2766.jpg
名称: 【美瓷陶瓷菜刀】美瓷(MYCERA)陶瓷菜刀 (黑色) AHG6.5B
链接: https://item.jd.com/1208647.html
相关度: 3.32669758797

图片: https://img12.360buyimg.com/n1/g5/M02/1B/0E/rBEDilAOl0UIAAAAALQwKP0D0YAAFXJgGoBCgAAtdY460.jpg
名称: 【乐瓷厨用刀】乐瓷LECI 6寸陶瓷厨用刀 LCQ4601A果绿
链接: https://item.jd.com/0654187.html
相关度: 2.65702748299

图片: https://img13.360buyimg.com/n1/g5/M02/1B/0E/rBEIDFA0LhMIAAAAAMTK4pTs\_MAAFXJ0K0vwMAAxND046.jpg
名称: 【乐瓷日式刀】乐瓷LECI 6寸陶瓷日式刀 LC5620A天蓝
链接: https://item.jd.com/0654123.html
相关度: 2.65702748299

图片: https://img12.360buyimg.com/n1/jfs/t568/105/1002536882/24732/798271d8/54a399b7Na5071618.jpg
名称: 【途耐陶瓷切片刀】途耐TONIFE 5英寸带刀套氧化陶瓷切片多用刀 抗菌免磨陶瓷刀具 紫色手柄
链接: https://item.jd.com/1308762.html
相关度: 2.41898393631

图片: https://img12.360buyimg.com/n1/jfs/t1747/10/105632526/138193/81b478/55cbf924N19394e71.jpg
名称: 【御碧陶瓷餐具套装】景德镇御碧陶瓷28头金粉世家餐具套装
链接: https://item.jd.com/1806547.html
相关度: 2.37011051178

图片: https://img13.360buyimg.com/n1/jfs/t1588/32/524669541/154563/98dbde28/55934ac6N1fd6c0e2.jpg
名称: 【斯凯绨骨瓷餐具套装】SKYTOP斯凯绨 陶瓷高档骨瓷餐具套装 40头金百合
链接: https://item.jd.com/0685123.html
相关度: 2.27745199203

图片: https://img10.360buyimg.com/n1/jfs/t4600/173/2079070138/220263/2d02d93/58eafc48N0c603ba9.jpg
名称: 【瓷状元花瓶摆件】瓷状元 天青釉鹿头尊 景德镇陶瓷器仿古青瓷花瓶玄关摆件电视柜装饰品 U2A004
链接: https://item.jd.com/4062785.html
相关度: 2.08135652542

图片: https://img13.360buyimg.com/n1/jfs/t10435/112/1282317558/84582/30dbfca1/59dec919N51928f46.jpg
名称: 【猪博士陶瓷保温灯泡】猪博士陶瓷加热灯 爬行动物宠物保温灯 养殖取暖灯泡小号150W (适合爬行类)
链接: https://item.jd.com/5604738.html
相关度: 1.97500717767
```

对图片搜索，显示图片地址、所在页地址、标题