

JavaScript 爬取网页并分析

任务分析:

- 1、爬取三个网站下的新闻数据，这里选择网易新闻网 (<https://news.163.com/>) ；
- 2、提取每条新闻的如下字段：标题，内容，发表日期，网址，关键词，作者，来源，评论等；
- 3、将爬取的数据写入数据库；
- 4、搭建前后端，实现对爬取数据的查询搜索分析等功能；

注：这篇博客只会对关键代码进行解析，完整代码在 GitHub 中

代码链接：<https://github.com/zgl-ai/a-crawler-about-javascript>

爬虫部分

首先是爬取网易新闻网 (<https://news.163.com/>)

引入一些必须的包

```
var fs = require('fs');
var myRequest = require('request')
var myCheerio = require('cheerio')
var myIconv = require('iconv-lite')
require('date-utils');
```

fs 负责文件读写，request 负责获得服务端发来的 html 响应，cheerio 负责解析 html 文件，incov-lite 负责 html 文件的编码格式的转换；

爬取网页的基本信息

```
var source_name = "网易新闻";
var myEncoding = "gbk";
var seedURL = 'https://news.163.com/';
```

要注意的是这个网页的编码格式为 GBK，之前我使用的是“utf-8”发现出现了乱码。

还有一些字段格式的声明在之后解释

爬虫的步骤一般如下：

- 1、向目标服务端发送一个种子页面请求；
 - 2、服务端返回这个页面的 html 文件；
 - 3、转码，解析文件，获得所需要的 tag 中信息；
 - 4、如果还需要获取种子页面下一些超链接的信息，那么就需要重复以上几步；
- 本次的实验也是按照以上的步骤进行的。

首先是构造相应的种子页面请求

```
//防止网站屏蔽我们的爬虫
var headers = {
```

```
'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_1)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/39.0.2171.65
Safari/537.36'
}

//request 模块异步 fetch url
function request(url, callback) {
var options = {
url: url,
encoding: null,
//proxy: 'http://x.x.x.x:8080',
headers: headers,
timeout: 10000 //
}
myRequest(options, callback)
}
```

这是基于 http 协议构造的请求信息，包括了头部信息，网址，时间限制等信息。

得到种子页面的 html 文件后，解码并且解析

```
var html = myIconv.decode(body, myEncoding);
//console.log(html);
//准备用 cheerio 解析 html
var $ = myCheerio.load(html, { decodeEntities: true });
```

我们的目的是通过网易新闻网的主页面，访问主页面中的新闻链接来获得新闻的信息。因此接下来我们要做的是，分析网易新闻网主页面中的有超链接，判断它是不是一个新闻的超链接（这里只提取文本类新闻）。

通过对比发现，网易所有文本类新闻的 URL 都有一个共同点：



那就是其中都有一个“article”字段。基于此来判断是不是一个有效的新闻链接
首先需要规范一下 url 的格式：

url 不能是“undefined”，意思就是不能为空，

```
if (typeof(href) == "undefined") { // 有些网页地址 undefined
return true;
}
```

url 必须是以大写或者小写的“http: //” 或者 “https://” 开头，否则就在前面加上 “http: ” 路径名规范为 https: //news.163.com/……

```
if (href.toLowerCase().indexOf('http://') >= 0 ||
href.toLowerCase().indexOf('https://') >= 0) {
myURL = href; //http://开头的或者 https://开头
//console.log(href);
}
else if (href.startsWith('//')) myURL = 'http:' + href; ////
开头的
else {
myURL = seedURL.substr(0, seedURL.lastIndexOf('/') + 1) +
href; //其他
//console.log(href);
}
```

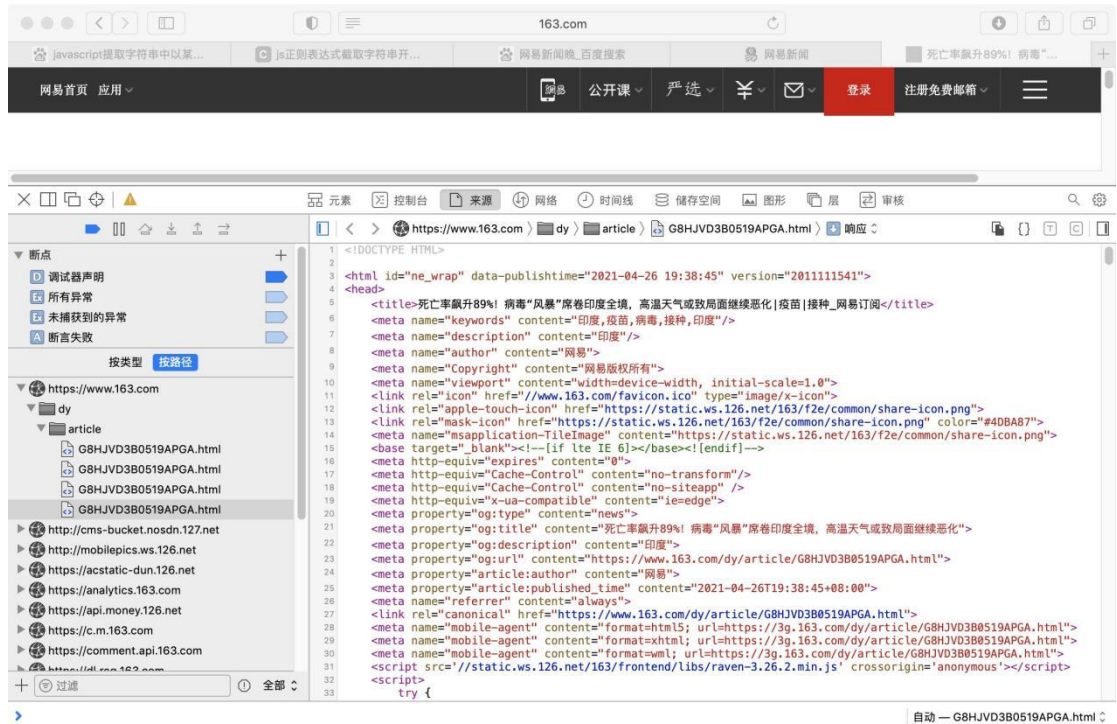
规范格式后就需要通过正则表达式的匹配来看有没有 article 字段:

```
var url_reg = /article/
```

```
if (!url_reg.test(myURL)){
//console.log(myURL);
return; //检验是否符合新闻 url 的正则表达式
}
```

接下来就是提取新闻的信息（标题，内容，发表日期，网址，关键词，作者，来源，评论等）；需要分析一下网页的源文件





发现标题存放位置，在<title>标签中

<title>死亡率飙升89%！病毒“风暴”席卷印度全境，高温天气或致局面继续恶化 | 疫苗 | 接种_网易订阅</title>

因此定义读取标题格式

```
var title_format = $("title").text();
```

关键词存储格式，在<meta>标签中的 keywords-content 键值对中

<meta name="keywords" content="印度,疫苗,病毒,接种,印度" />

```
var keywords_format = "  
$('meta[name=\"keywords\"]').eq(0).attr(\"content\");
```

正文内容是在，“post_body”这一个类别中

```
<div class="post_body">  
<p> 本文来自：时代周报 作者：刘沐轩</p>  
<p> 随着第二波疫情的暴发，印度在4月25日再次刷新了全球新冠病毒单日确诊人数最高纪录，一天内新增确诊超35万人，累计确诊人数达1731万人。仅在过去一周，印度就新增确诊病例225万，死亡率飙升89%。《纽约时报》甚至认为，有证据显示，真实死亡人数要远高于官方报告。</p>  
<p> 印度总理莫迪在4月25日公开承认，病毒感染的“风暴”正在席卷印度全境。他敦促印度居民尽快接种疫苗。</p>  
<br/></p>  
<p> 在印度新德里的大火葬场，逝者大多数为新冠肺炎疫情的受害者。（图源：印度时报）</p>  
<p> 但印度新冠肺炎疫苗库存紧张已经是人尽皆知的事实，即使是全球最大的疫苗生产国，但因资金和原材料问题，目前印度疫苗的产能仍难以提升。如果没有外部援助，印度很可能深陷在疫情泥潭中，不仅经济复苏无望，还会破坏全球供应链。</p>  
<p> 继中国、英国、法国和德国之后，美国也在一片批评声中转变了态度，同意伸出援手。当地时间4月25日，美国宣布将立即为印度提供生产新冠肺炎疫苗所需的原材料。尽管如此，美国仍然打算动用本国多余的数百万剂疫苗库存。</p>  
<p> 讽刺的是，美国国家安全顾问沙利文呼吁，“作为全世界新冠确诊病例最多的两个国家，美国与印度应该保持团结。”</p>  
<strong>恶性循环引爆印度疫情</strong></p>  
<p> 从今年3月初印度政府对疫情盲目乐观，到4月初完全没有社交限制的大规模集会和宗教节日，印度疫情在暴发后陷入了恶性循环中。</p>  
<p> 由于一直没有采取有力措施彻底控制疫情，病毒在印度的传播中也不断进化。继发现高传染性的双重变异新冠病毒B.1.617之后，据《印度时报》报道，印度国家生物医学基因组学研究所研究员坎特瓦在4月25日表示，印度近期还新发现了一种三重变异的新病毒株。这种毒株很可能使得接种过疫苗的人再次感染。</p>  
<br/></p>  
<p> 由于没有多余的病床，一名呼吸困难的印度患者在车内接上呼吸机。（图源：路透社）</p>  
<p> 值得注意的是，除了社交隔离措施松懈，炎热的天气也加剧了印度这波疫情的传播。由多名印度学者联合发表的研究报告指出，气温升高有利于新冠肺炎病毒的传播。他们预测，随着印度各地气温升高，病例还将进一步增加。而近期印度境内多地也出现了持续的高温天气，新德里在未来一周的最高温度将达到43度。</p>  
<p> 在疫苗供应方面，由于印度政府此前对国内疫情过于乐观，已经将大量印度仿制疫苗出口。不仅如此，由于后续资金也不足，导致印度血清研究所等国内的疫苗生产难以提高产能。</p>  
<p> 居住在美国泰尔米纳德和金奈市的一位中资企业员工胡先生在去年4月26日对时代周报记者表示，“印度有疫苗生产设施，但是没有原料。原料大多来自中美、美国之前不出口，印度也不太愿意从中国大量进口。”</p>  
<p> 这一切最终导致了印度医疗体系的崩溃，进一步抬高了疫情的死亡率。由于病床和氧气短缺，目前印度各地的医院已开始拒收病患。</p>  
<p> 据《印度快报》报道，在新德里一家医院的重症监护室里，一晚上至少有20名新冠患者因“氧气压力太低”而逝世。</p>  
<p> 对于印度疫情，复旦大学附属华山医院感染科主
```

```
var content_format = $(".post_body").text();
```

同理其他的所有的读取格式都通过对比网页源码来获得

```
var keywords_format = "  
$('meta[name=\"keywords\"]').eq(0).attr(\"content\");
```

```

var title_format = "$('title').text()";
//<meta property="article:published_time"
content="2021-04-26T14:47:04+08:00">
var date_format = "
$('meta[property=\\\"article:published_time\\\"]').eq(0).attr(\\\"content\\\")";
//<meta name="author" content="网易">
var author_format = "
$('meta[name=\\\"author\\\"]').eq(0).attr(\\\"content\\\")";
var content_format = "$('.post_body').text()";
var desc_format = "
$('meta[name=\\\"description\\\"]').eq(0).attr(\\\"content\\\")";
var source_format = "$('.post_info').text()";

```

将目标信息通过 evel 读取并且存入 fetch 结构体中

```

var fetch = {};
fetch.title = "";
fetch.content = "";
fetch.publish_date = (new Date()).toFormat("YYYY-MM-DD");
//fetch.html = myhtml;
fetch.url = myURL;
fetch.source_name = source_name;
fetch.source_encoding = myEncoding; //编码
fetch.crawltime = new Date();

```

发表日期的写入的时候要注意

因为有的文章没有发表日期,所以需要填上“未知”,再将所有的日期转换为“YYYY-MM-DD”的格式

```

if (date_format != "") {
fetch.publish_date = eval(date_format); //刊登日期
if (!fetch.publish_date){
fetch.publish_date = "未知";
}else{
//fetch.publish_date = regExp.exec(fetch.publish_date)[0];
fetch.publish_date = fetch.publish_date.replace('年', '-')
fetch.publish_date = fetch.publish_date.replace('月', '-')
fetch.publish_date = fetch.publish_date.replace('日', '')
fetch.publish_date = new
Date(fetch.publish_date).toFormat("YYYY-MM-DD");
}
}

```



```
}  
}
```

来源部分也需要通过正则表达式来提取

```
if (source_format == "") fetch.source = fetch.source_name;  
else {  
    fetch.source = eval(source_format).replace("\r\n", ""); //  
    来源  
    var matchReg = /(?!<=来源: ).*?(?=.\n)/gi;  
    fetch.source=(fetch.source.match(matchReg));  
    if(fetch.source==null){  
        fetch.source=source_name;  
    }  
}
```

完成上面所有的步骤，就可以提取新闻的信息了，一下选取一篇新闻提取的信息来展示

```
{  
    "title": "临危受命！超30人被拿下的当天上午，湖北女将跨省救火 | 鄂钢|昆钢|钢铁|钢铁厂",  
    "content": "\n                撰文 | 余晖    政知君注意到，惊现窝案的昆钢",  
    "publish_date": "2021-04-16",  
    "url": "https://www.163.com/dy/article/G7LPG08G051482MP.html",  
    "source_name": "网易新闻",  
    "source_encoding": "gbk",  
    "crawltime": "2021-04-27T04:16:53.924Z",  
    "keywords": "湖北,鄂钢,昆钢,钢铁,钢铁厂",  
    "author": "网易",  
    "source": [  
        "政知新媒"  
    ],  
    "desc": "临危受命！超30人被拿下的当天上午，湖北女将跨省救火    ,湖北,鄂钢,昆钢,钢铁,钢",  
}
```

将爬取到的所有信息存储到 mysql 数据库中

这需要已经提前下载好的 mysql 数据库,并且创建好 crawl 库

```
create database crawl;
```

```
use crawl;
```

创建表 fetches

```
CREATE TABLE `fetches` (  
    `id_fetches` int(11) NOT NULL AUTO_INCREMENT,  
    `url` varchar(200) DEFAULT NULL,  
    `source_name` varchar(200) DEFAULT NULL,  
    `source_encoding` varchar(45) DEFAULT NULL,  
    `title` varchar(200) DEFAULT NULL,
```

```

`keywords` varchar(200) DEFAULT NULL,
`author` varchar(200) DEFAULT NULL,
`publish_date` date DEFAULT NULL,
`crawltime` datetime DEFAULT NULL,
`content` longtext,
`createtime` datetime DEFAULT CURRENT_TIMESTAMP,
PRIMARY KEY (`id_fetches`),
UNIQUE KEY `id_fetches_UNIQUE` (`id_fetches`),
UNIQUE KEY `url_UNIQUE` (`url`)
) ENGINE=InnoDB DEFAULT CHARSET=gbk;

```

JavaScript 连接数据库需要相应的包

```
var mysql = require('./mysql.js');
```

插入数据库的语句，和插入的变量

```

var fetchAddSql = 'INSERT INTO
fetches(url,source_name,source_encoding,title,' +
'keywords,author,publish_date,crawltime,content)
VALUES(?,?,?,?,?,?,?,?)';
var fetchAddSql_Params = [fetch.url, fetch.source_name,
fetch.source_encoding,
fetch.title, fetch.keywords, fetch.author,
fetch.publish_date,
fetch.crawltime.toFormat("YYYY-MM-DD HH24:MI:SS"),
fetch.content
];

```

执行插入语句

```

//执行 sql，数据库中 fetch 表里的 url 属性是 unique 的，不会把重复的
url 内容写入数据库
mysql.query(fetchAddSql, fetchAddSql_Params, function(qerr,
vals, fields) {
if (qerr) {
console.log(qerr);
}
}); //mysql 写入

```

至此，从网易新闻官网爬取数据，并且存入数据库的任务已经完成，数据库里的数据展示如下

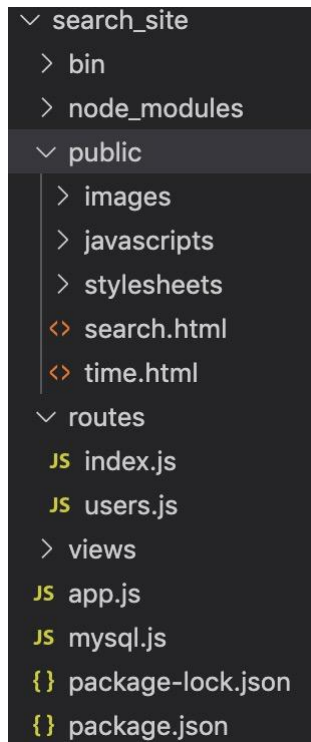
id_fetches	url	source_name	source_encoding	title	keywords	author
1	https://www.163.com/dy/article/...	网易新闻	gbk	曹峰：“台美新交”43年来，首位美大使访...	台军,美军,演习,帕劳,关岛	网易
2	https://www.163.com/dy/article/...	网易新闻	gbk	美媒：对新疆的“种族灭绝”指控纯属子虚...	种族灭绝,指控,英国政府,拜登,恐怖主义,...	网易
3	https://www.163.com/dy/article/...	网易新闻	gbk	基辛格：中国的复苏不令人惊讶，美国应对...	基辛格,美国,中国,国务院,拜登,外交政策,...	网易
4	https://www.163.com/dy/article/...	网易新闻	gbk	俄罗斯做了什么？让捷克改口，乌克兰悔悟...	纳瓦利内,乌克兰,俄罗斯,捷克,普京,泽连...	网易
5	https://www.163.com/dy/article/...	网易新闻	gbk	中国使馆发推宣布“800台制氧机运抵德里”...	印度,制氧机,德里,近邻,远亲,莫迪,印度	网易
6	https://www.163.com/dy/article/...	网易新闻	gbk	刘宗义：失控的疫情，正在侵蚀印度中产以...	印度,刘宗义,莫迪,印度政府,疫苗,印度疫情	网易
7	https://www.163.com/dy/article/...	网易新闻	gbk	苹果发布 watchOS 7.4 正式版：支持解...	iphone,apple,ios,苹果,watchos	网易
8	https://www.163.com/dy/article/...	网易新闻	gbk	台陆委会：恢复大陆商务人士赴台申请，再...	陆委会,国台办,九二共识,台湾当局,邱太三	网易
9	https://www.163.com/dy/article/...	网易新闻	gbk	民航西南空管局完成绵阳机场飞行程序设计...	飞行,绵阳,机场,民航,空管局	网易
10	https://www.163.com/dy/article/...	网易新闻	gbk	2021年北京西城小学入学24问，权威解答...	入学,户籍,户口,居住证	网易
11	https://www.163.com/dy/article/...	网易新闻	gbk	中国驻印度大使：边界争议是客观现实，包...	中印关系,中印,外交,印度大使	网易
12	https://www.163.com/dy/article/...	网易新闻	gbk	疫情海啸吞没印度，宝莱坞明星却孔雀开...	马尔代夫,宝莱坞,印度,海啸,伊朗,宝莱坞...	网易
13	https://www.163.com/dy/article/...	网易新闻	gbk	全新奔驰GLC内饰谍照首曝，中控台布局...	奔驰glc,奔驰c级,新车,内饰,谍照,奔驰	网易
14	https://www.163.com/dy/article/...	网易新闻	gbk	赵丽颖带火的“衬衫”，专治身高矮！155c...	衬衫,牛仔裙,减龄,长袖衬衫,半身裙,斑点	网易
15	https://www.163.com/dy/article/...	网易新闻	gbk	美团王兴越“界”招“祸”！饿了么 外卖 大...	王兴,美团,饿了么,外卖,大众点评,互联网公	网易
16	https://www.163.com/dy/article/...	网易新闻	gbk	Brembo发布最新制动卡钳及线控制动技术...	制动器,brembo,卡钳,brake	网易
17	https://www.163.com/dy/article/...	网易新闻	gbk	北京警方2020年以来铲除制假窝点3...制假窝点,黑窝点,窝点,假酒,刑事拘留		网易
18	https://www.163.com/dy/article/...	网易新闻	gbk	2吨台湾菠萝运抵澳大利亚 每箱售价高达9...	澳大利亚,台湾,菠萝,运抵,农委会	网易
19	https://www.163.com/dy/article/...	网易新闻	gbk	番禺出手：禁止假人才炒房 新盘不接受指...	番禺,炒房,调控,指导价,预售证,人才购房	网易
20	https://www.163.com/dy/article/...	网易新闻	gbk	20个涉嫌渎职官员被拉清单！她忏悔鞠躬...	鄂尔多斯市,渎职官员,德林郭勒盟,内蒙古...	网易
21	https://www.163.com/dy/article/...	网易新闻	gbk	重庆首个航空运动旅游营地开营首飞 飞机 ...	旅游,航空,飞机,飞行,空域	网易
22	https://www.163.com/dy/article/...	网易新闻	gbk	俄罗斯“一箭36星”在俄东方航天发射场成...	卫星,发射场,俄罗斯,航天,推进器	网易
23	https://www.163.com/dy/article/...	网易新闻	gbk	新入列海南舰最新训练画面曝光。“提灯饼...	舰载机,辽宁舰,舰艇,山东舰,直升机,中国海	网易
24	https://www.163.com/dy/article/...	网易新闻	gbk	西班牙13岁索菲亚公主个子超170了！大公...	索菲亚,zara,泡泡袖,内搭,裙子,大衣	网易
25	https://www.163.com/dy/article/...	网易新闻	gbk			

同理还爬取了搜狐（<https://www.sohu.com>）、腾讯新闻（<https://www.qq.com>）的新闻。总共的数据有三百多条。工作量不大，就是需要向上文那样，产源码的结构，找出目标字段所在的标签，然后提取出来存入数据库。因此写了三个爬虫文件分别爬取这三个代码

```
JS crawl.js
JS crawl2.js
JS crawl3.js
```

搜索和分析

这里需要一个前后端交互的功能, 用到了 **express** 脚手架, 具体的搭建方法这里就不详述了。最后可以创建一个路由架构。



主要的工作量是 search.html 文件，time.html 文件和 index.js 文件。

search.html 文件：实验按标题搜索关键词的前端网页。

首先是需要一个表单，来接受用户输入的关键词并传输给后端

```
<body>
<form>
<br> 标题: <input type="text" name="title_text">
<input class="form-submit" type="button" value="查询">
</form>
```

标题:

用一个表来展示数据库传来的数据

```
<div class="cardLayout" style="margin: 10px 0px">
<table width="100%" id="record2"></table>
</div>
```

接下来需要通过 javascript 语言来与后端交互数据

与后段交互数据:

构造路由与出入数据

```
$va='/process_get?title=' + $("input:text").val();
```

传入数据后需要接受后端传来的数据，在 data 这个变量中

```
$.get($va, function(data) {
```

解析 data 数据存入表 record2 中

```
$("#record2").empty();
```

```

$("#record2").append('<tr
class="cardLayout"><td>url</td><td>source_name</td>' +
'<td>title</td><td>author</td><td>publish_date</td></tr>');
;
for (let list of data) {
let table = '<tr class="cardLayout"><td>';
$i=0;
Object.values(list).forEach(element => {
if($i==0){
table += ('<a href='+element+'>' + element + '</a>' +
'</td><td>');
}else{
table += (element + '</td><td>');
}
$i++;
});
$("#record2").append(table + '</td></tr>');

```

这里是一列一列的读取 data，然后对每列再一个一个元素的读取。存入的时候需要了解 html 文件中表格的格式

```

<table border="1">
<tr>
<td>row 1, cell 1</td>
<td>row 1, cell 2</td>
</tr>
<tr>
<td>row 2, cell 1</td>
<td>row 2, cell 2</td>
</tr>
</table>

```

除此之外，我还把 url 变成超链接的形式。url 位于每一列的第一个元素中。上图的红框所示。

引入一个超链接指向 time.html

```

<a href='time.html'>点击进行时间热度分析</a>

```

index.js:负责后端的，与数据库交互，与前端交互

首先是 search.html 路由到的函数

```

router.get('/process_get', function(request, response) {
//sql 字符串和参数
var fetchSql = "select
url,source_name,title,author,publish_date " +

```

```

"from fetches where title like '%" + request.query.title +
"%';
mysql.query(fetchSql, function(err, result, fields) {
response.writeHead(200, {
"Content-Type": "application/json"
});
response.write(JSON.stringify(result));
response.end();
});
});

```

收到前端传来的 title 数据之后，向数据库发送查询命令

Select url,source_name,title,author,publish_date

From fetches

Where title like ‘%……%’（这个是查询包含关键词\$title 的语法）

查询之后，结果写到了 result 变量中，然后以 json 格式传到前端的数据中。

这两个 part 运行的结果如下：

工作目录下运行 node bin/www

浏览器输入：localhost: 3000/search.html

标题: 查询

点击进行时间热度分析

url	source_name	title	author	publish_date
https://www.163.com/dy/article/G8HVOHRJ0518DCRR.html	网易新闻	鞋子最能表达品位，你的鞋选对了吗？ 穆勒鞋高跟鞋显高上脚_网易订阅	网易	2021-04-25T16:00:00.000Z
https://www.163.com/dy/article/G8JLBDUB05525YPK.html	网易新闻	超冲不光你特斯拉有：小鹏充电桩将在所有地级市进行布局 电动汽车_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G8GA3TTE0515CLPL.html	网易新闻	嫦娥六号任务预计2024年前后实施 或将继续月背征途 南极月球着陆点_网易订阅	网易	2021-04-25T16:00:00.000Z
https://www.163.com/dy/article/G69H8AF40543B4S9.html	网易新闻	警惕！“台美斯交”43年来，首位美大使访台，蔡当局不再遮掩？ 美军演习 帕劳_网易订阅	网易	2021-03-28T16:00:00.000Z
https://www.163.com/dy/article/G8JJ4OSG05504DPG.html	网易新闻	赵立坚发推：如果原作者还活着... 日本政府神奈川冲浪里 插画师_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G7LFIRT20514R9L4.html	网易新闻	俄罗斯回应美国新制裁：美方会为两国关系恶化付出代价 俄外交部_网易订阅	网易	2021-04-14T16:00:00.000Z
https://www.163.com/dy/article/G8KDDU1J0529AQIE.html	网易新闻	锡伯社：兰德尔几乎每个回合被包夹 但他是球队的发动机 纽约尼克斯队_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G8CACDFD0538A2QY.html	网易新闻	虽然贾静雯已经是妈妈辈 但她穿蓝裙真高级！ 开叉 裙子 长裙 v领_网易订阅	网易	2021-04-23T16:00:00.000Z
https://www.163.com/dy/article/G8JVAFT0547DDEC.html	网易新闻	丰田全新兰德酷路泽实拍！搭3.5T V6引擎，有望8月上市 新车 发动机 雷克萨斯_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G8FEU75H050538A2QY.html	网易新闻	谁说针织裙不时髦？这样穿好看又百搭 穿搭 裙子 紧身 版型_网易订阅	网易	2021-04-24T16:00:00.000Z
https://www.163.com/dy/article/G8HBQP600521QBNK.html	网易新闻	绝密线人躲过层层封锁，从緬北带出一份105人名单！ 杨南 南昌市公安局 关尾_网易订阅	网易	2021-04-25T16:00:00.000Z
https://www.163.com/dy/article/G8JJGJ790527JOC.html	网易新闻	颜值再升级 疑似新款大众威然实车曝光 新车 内饰 腰线 实车图_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G8KD4OB0529AQIE.html	网易新闻	比尔生涯砍下40+的比赛7胜21负 胜率25%为历史最低 马刺队 奇才_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G8EPMFF0540RJBN.html	网易新闻	这些少女感的穿搭，让你精致感满满 纱裙 背带裙 少女风 裙装 连衣裙 牛仔_网易订阅	网易	2021-04-25T16:00:00.000Z

time.html 文件：实现按照时间热度搜索关键词的前端页面。逻辑和 search.html 差不多

inde.js 中负责处理 time.html 文件的函数

```

router.get('/time_get', function(request, response) {
//sql 字符串和参数
var fetchSql = "select publish_date,COUNT(keywords) " +

```

```

"from fetches where keywords like '%" + request.query.title
+ '%" GROUP BY publish_date ORDER BY COUNT(keywords) DESC";
mysql.query(fetchSql, function(err, result, fields) {
response.writeHead(200, {
"Content-Type": "application/json"
});
console.log(result);

response.write(JSON.stringify(result));
response.end();
});
});

```

它在接受到 title 之后，向数据库发送查询指令

Select publish_date,count(keywords)

From fetches

Where keywords like ‘%……%’

Group by publish_date

Order by count(keywords) desc;

这样可以做到使搜索结果按照发表日期做好统计。

点击这个之后就可以看到运行结果

标题: <input type="text"/>	查询
点击进行时间热度分析	

url	source_name	title	author	publish_date
https://www.163.com/dy/article/G8HVQHRJ0518DCRR.html	网易新闻	鞋子最能表达品位，你的鞋选对了吗？ 穆勒鞋 高跟鞋 显高 上脚_网易订阅	网易	2021-04-25T16:00:00.000Z
https://www.163.com/dy/article/G8JLBDUB0525YYPK.html	网易新闻	超冲不光你特斯拉有：小鹏充电桩将在所有地级市进行布局 充电 电动汽车_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G8GA3ITE0515CLPL.html	网易新闻	嫦娥六号任务预计2024年前后实施 或将继续月背征途 南极 月球 着陆点 月球背面_网易订阅	网易	2021-04-25T16:00:00.000Z
https://www.163.com/dy/article/G69H8AF40543B4S9.html	网易新闻	警惕！“台美断交”43年来，首位美大使访台，蔡当局不再遮掩？ 美军 演习 帕劳 关岛_网易订阅	网易	2021-03-28T16:00:00.000Z
https://www.163.com/dy/article/G8JL4OSG05504DPG.html	网易新闻	赵立坚发推：如果原作者还活着... 日本政府 神奈川 冲浪 插画师_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G7JLPIRT20514R9L4.html	网易新闻	俄罗斯回应美国新制裁：美方会为两国关系恶化付出代价 俄外交部_网易订阅	网易	2021-04-14T16:00:00.000Z
https://www.163.com/dy/article/G8KDDU1J0529AQJE.html	网易新闻	锡伯社：兰德尔几乎每个回合被包夹 但他是球队的发动机 纽约尼克斯队_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G8CACDFD0538A2OY.html	网易新闻	虽然贾静雯已经是妈妈辈 但她穿蓝裙真高级！ 开叉 裙子 长裙 v领_网易订阅	网易	2021-04-23T16:00:00.000Z
https://www.163.com/dy/article/G8JJVAF70547DDEC.html	网易新闻	丰田全新兰德酷路泽实拍！搭3.5T V6引擎，有望8月上市 新车 发动机 雷克萨斯 s_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G8EU75H0538A2OY.html	网易新闻	谁说针织裙不时髦？这样穿好看又百搭 穿搭 裙子 紧身 版型_网易订阅	网易	2021-04-24T16:00:00.000Z
https://www.163.com/dy/article/G8HBQP600521OBNK.html	网易新闻	绝密线人躲过层层封锁，从緬北带出一份105人名单！ 杨雨 南昌市公安局 关尾_网易订阅	网易	2021-04-25T16:00:00.000Z
https://www.163.com/dy/article/G8JJGJ790527JOC.html	网易新闻	颜值再升级 疑似新款大众威然实车曝光 新车 内饰 腰线 实车图_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G8KDDU1J0529AQJE.html	网易新闻	比尔生涯砍下40+的比赛7胜21负 胜率25%为历史最低 马刺队 奇才_网易订阅	网易	2021-04-26T16:00:00.000Z
https://www.163.com/dy/article/G8EPMFFJ0540RJBN.html	网易新闻	这些少女感的穿搭，让你精致感满满 纱裙 背带裙 少女风 裙装 连衣裙 牛仔_网易订阅	网易	2021-04-25T16:00:00.000Z

localhost	
腾讯网_百度搜索	html表格_百度搜索
HTML 表格	localhost:3000/time.html

搜索要查询的关键词:	<input type="text"/>	查询
时间		数量
2021-04-27T16:00:00.000Z		245
2021-04-26T16:00:00.000Z		65
2021-04-25T16:00:00.000Z		22
2021-04-24T16:00:00.000Z		11
2021-04-19T16:00:00.000Z		7
2021-04-23T16:00:00.000Z		3
2021-03-28T16:00:00.000Z		1
2021-04-14T16:00:00.000Z		1
2021-04-22T16:00:00.000Z		1
2021-04-10T16:00:00.000Z		1
2021-03-13T16:00:00.000Z		1

发现目前统计的新闻数据中，4 月 27 发表的新闻最多。

搜索“裙子”关键词

搜索要查询的关键词:	<input type="text" value="裙子"/>	查询
时间		数量
2021-04-23T16:00:00.000Z		1
2021-04-24T16:00:00.000Z		1
2021-04-22T16:00:00.000Z		1
2021-04-25T16:00:00.000Z		1

总结

这次项目的几个难点

1、为了找到新闻 url，需要查看很多新闻的 url，找到相似处，由于每个企业对网页的格式定义不同，我们也需要相应的改写相关的正则表达式；

2、读取相关字段，如标题，日期作者等等。在选取本次实验的三个种子网页（腾讯，网易，搜狐）之前，我还尝试爬取过虎扑，东方财富等其他新闻网页。但是在爬取过程中总是碰到一些奇怪的问题。比如有的字段藏在了很复杂的标签块中，在爬虫代码中不好以统一的格式定义，因此在代码运行的过程中总是出现爬取到空字段的情况。还有就是每个网站的编码格式不一样，有的是 utf-8 有的又是 gbk，所以在爬取的时候需要都试一试。

虽然如此，这些网页还是会采用一些通用的标准来编写，比如“标题，关键字，摘要”等信息。

3、统一每个网页爬取到的信息。在爬取“时间”这一个字段是，有的网页是“2021-4-29”，有的又是“4/29/2021”，有的又是“2021 年 4 月 29 号”，虽然统一起来不是很难吧，但也很麻烦。

4、前后端路由。由于对 Javascript 语法不是很熟悉，所以在编写代码中也遇到很多问题，就比如前后端路由的时候。具体一点，就是前端怎么向后端传数据，后端接受数据后怎么读取出来。后端怎么向前端传数据，前端接收到之后，怎么将 json 格式的数据读取出来。在搞清语法上花了很多时间。

5、本来之前还想做很多任务的，比如对展示结果进行分页，用 css 做一个好看一点的前端，在交互上做的好一点。但是自己的这方面知识确实太贫瘠了，作业提交时间也快到了。

因此也就完成了这些基本功能。

好在之前有数据库，python flask, php 等相关基础，因此还是完成了这次实验。