

# Node.js新闻网站爬虫、并实现爬取结果查询

## 新闻网站爬取

本项目爬取的网站为**东方财富和财经网**，集中在4.27日-4.29日三天的傍晚进行网页爬取，总共爬取了649条有效网站，并将新闻网站内容进行解析，选取了网站链接、新闻标题、关键词、网站来源、摘要、发布时间、爬取时间、新闻内容存储于**Mysql**数据库中。

### 引用的模块

#### request模块

该模块在读取种子页面，遍历新闻链接的时候和读取新闻链接的时候时使用。该模块使http请求变得更简单。

#### iconv-lite模块

该模块的作用就是转码，用iconv-lite.decode()转码，可以设置成'utf-8'，以规避乱码产生。

#### cheerio模块

该模块是爬虫中很重要的一个模块，其主要作用是加载你要访问的HTML页面，即把HTML页面翻译给处理器。有了它我们才能对网页做进一步处理操作。

### 分析新闻页面

下方部分代码和图片是我对东方财富网网页内容进行解析的方式（全部代码可以到我的github项目中去看），首先要进入东方财富的主页（种子页面）读取所有新闻链接（子网页）这一过程需要使用正则表达式来帮助判断新闻链接是否合法。之后根据网页的源代码（源代码查看方式：Fn+F12）分析所有新闻页面，解析出结构化数据。

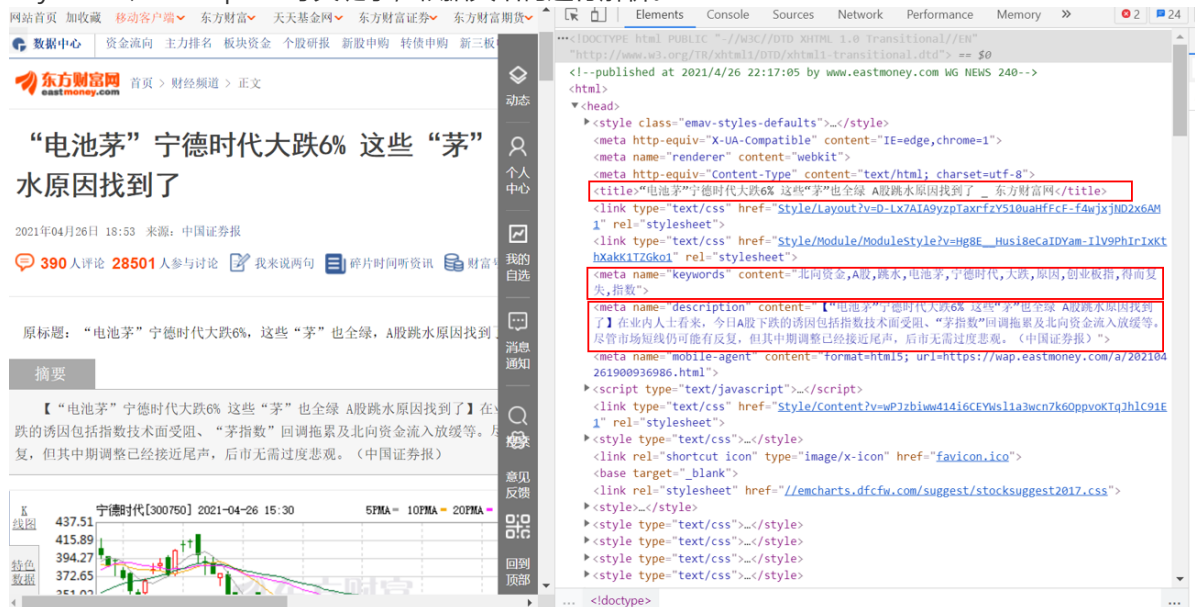
以下代码位于crawl\_eastmoney.js文件中，是对东方财富网进行的爬取。而对财经网进行爬取的代码是在crawl\_caijing.js中。对财经网的爬取与对东方财富网的爬取很类似，只是在网页链接的匹配和关键字的匹配上有所不同，在此就不展示了，具体可以看我的代码。create\_table.sql主要是为了在数据库中创建数据表，mysql.js是为了在node.js中连接到数据库，这两个文件在后续会进行说明。

```
// $表示查找，如$('a')查找以a开头的子网页链接。'#'表示查找的是id，'.'表示查找的是class
var seedURL_format = $('a');
var keywords_format = $('meta[name="keywords"]').eq(0).attr("content");
//eq(0)是从头开始把content的数据都取出来
var description_format =
$('meta[name="description"]').eq(0).attr("content");
var title_format = $('title').text();
var date_format = $('time').text();
var content_format = $('body').text();

//使用正则表达式选取合适的网页链接
var url_reg = /\https?:\/\/finance[.]eastmoney[.]com\/a\/\d*.html/;
//使用正则表达式选取网页发表时间
var regExp = /((\d{4}|\d{2})(-|\/|\.)\d{1,2}\3\d{1,2})|(\d{4}年\d{1,2}月\d{1,2}日)/
```

东方财富网站主网页（种子网页）上有很多链接，但是并非所有的链接都为新闻链接，也有一些链接是指向其他非新闻页面的。而我们的目的是找出所有新闻页面，所以需要对URL进行选择，通过观察发现，在东方财富网站中，所有新闻网站都是以<http://finance.eastmoney.com/a/...html>的结构构成，所以根据此构造正则表达式来匹配这样的链接，从而筛掉其他非新闻链接。

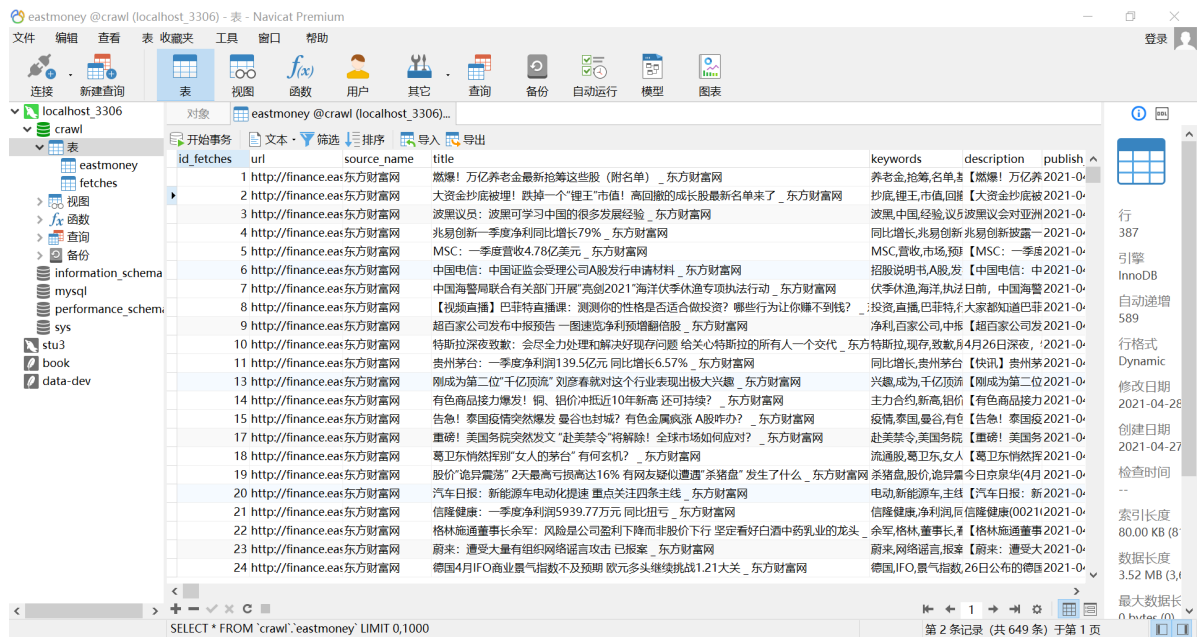
然后在东方财富网主页并随机选择一个新闻网页进入其中，在控制台查看其源代码，找到title、keywords、description等关键字，根据其结构进行解析。



```
console.log(qerr);
}
}); //mysql写入
```

```
//查询当前的URL是否已经在mysql数据库中
var fetch_url_Sql = 'select url from eastmoney where url=?';
var fetch_url_Sql_Params = [myURL];
mysql.query(fetch_url_Sql, fetch_url_Sql_Params, function(qerr, vals, fields) {
    if (vals.length > 0) {
        console.log('URL duplicate!')
    } else newsGet(myURL); //读取新闻页面
});
```

把数据存入数据库之后可以通过Navicat查看mysql数据库中的内容，如果有乱码的话，可以修改解码方式，具体为utf-8还是GBK根据网页的不同而各异。我们可以在图片的右下角看到该表格中共有649条数据，也没有乱码，由此可见数据存储到数据库是成功的。



id_fetches	url	source_name	title	keywords	description	publish
1	http://finance.eas	东方财富网	燃爆！万亿养老金最新抢筹这些股（附名单）	养老金,抢筹,名单,【燃爆！万亿养老金2021-0-		
2	http://finance.eas	东方财富网	大资金抄底被埋！跌停一个“锂王”市值！高回报的成长股最新名单来了	抄底,锂王,市值,回【大资金抄底被 2021-0-		
3	http://finance.eas	东方财富网	波黑议员：波黑可学习中国的很多发展经验	波黑,中国,经验,议【波黑议会 2021-0-		
4	http://finance.eas	东方财富网	兆易创新一季度净利润同比增长79%	兆易创新,【兆易创新创新披露—2021-0-		
5	http://finance.eas	东方财富网	MSCI：一季度营收4.78亿美元	MSCI,营收,市场,预【MSCI：一季度 2021-0-		
6	http://finance.eas	东方财富网	中国电信：中国证监会受理公司A股发行申请材料	招股说明书,A股,发【中国电信：中 2021-0-		
7	http://finance.eas	东方财富网	中国海警局联合有关部门开展“利剑2021”海洋伏季休渔专项执法行动	伏季休渔,海洋,执【中国海警 2021-0-		
8	http://finance.eas	东方财富网	【视频直播】巴菲特直播课：测试你的性格是否适合做投资？哪些行为让你赚不到钱？	投资,直播,巴菲【大家都知 2021-0-		
9	http://finance.eas	东方财富网	超百家公司发布中报预告 一图速览净利润预增翻倍股	净利润,百家公【超百家公 2021-0-		
10	http://finance.eas	东方财富网	特斯拉深夜致歉：会尽全力处理和解决好现存问题 给关心特斯拉的所有人一个交代	特斯拉,深夜,致【特斯拉, 2021-0-		
11	http://finance.eas	东方财富网	贵州茅台：一季度净利润139.5亿元 同比增长6.57%	贵州茅台,【快讯】贵州茅 2021-0-		
12	http://finance.eas	东方财富网	刚成为第二位“千亿顶流”刘彦春就对这个行业表现出极大兴趣	刘彦春,【刚成为第二位 2021-0-		
13	http://finance.eas	东方财富网	有色商品接力爆发！铜、铝价冲抵近10年新高 还可持续？	主力合约,新高,【有色商品 2021-0-		
14	http://finance.eas	东方财富网	告急！泰国疫情突然爆发 曼谷也封城？有色金属疯涨 A股咋办？	疫情,泰国,曼谷,【告急！泰 2021-0-		
15	http://finance.eas	东方财富网	重磅！美国务院突然发文“赴美禁令”将解除！全球市场如何应对？	赴美禁令,美国【重磅！美国 2021-0-		
16	http://finance.eas	东方财富网	葛卫东悄然挥别“女人的茅台” 有何玄机？	流通股,葛卫东,【葛卫东 2021-0-		
17	http://finance.eas	东方财富网	股价“诡异震荡”2天最高亏损高达16% 有网友疑似遭遇“杀猪盘” 发生了什么	杀猪盘,股价,诡【杀猪盘 2021-0-		
18	http://finance.eas	东方财富网	汽车日报：新能源车电动化提速 重点关注四条主线	电动,新能源车,主【汽车日报： 2021-0-		
19	http://finance.eas	东方财富网	信隆健康：一季度净利润5939.77万元 同比扭亏	信隆健康,净利【信隆健康 2021-0-		
20	http://finance.eas	东方财富网	格林施通董事长余军：风险是公司盈利下降而非股价下行 坚定看好白酒中药乳业龙头	余军,格林,董【格林施通 2021-0-		
21	http://finance.eas	东方财富网	蔚来：遭受大量有组织网络谣言攻击 已报案	蔚来,网络谣【蔚来：遭 2021-0-		
22	http://finance.eas	东方财富网	德国4月IFO商业景气指数不及预期 欧元多头继续挑战1.21大关	德国,IFO,景气【德国,IFO, 2021-0-		

## 爬取结果查询与展示

此处的查询过程为：

- 在网页中输入/选择所需查询关键字并提交
- 网页通过get方法传递参数给HTML文件，HTML文件进行解析并将参数交给node.js
- node.js连接到后端的mysql数据库进行语句查询
- node.js将查询结果返回给HTML文件，HTML以表格的形式展示到网页当中，并使用css进行美化。

爬取结果的查询其实可以直接在数据库中进行查询，但是考虑到在mysql数据库中直接查询需要会sql语句才能进行，不便于操作也不美观，所以本项目使用express脚手架来创建一个网站框架，具体框架如下所示。可以根据用户在网页中输入/选择的标题和时间进行新闻查询，并且能够将查询结果分页显示，此外还使用了css对网页进行了一定程度的美化。其中，标题的查询是只要包含查询词就会予以显示，不必输入完整的标题。标题和时间都并非必填项，如果用户不输入该项，则会放弃对该项进行筛选，如果所有查询项都不输入，则会返回数据库中所有结果。

bin	
www	启动文件
public	静态资源文件目录，该文件夹下不需要映射可以直接访问
search.html	功能为查询爬取结果
time.html	功能为时间热度分析
style.css	美化页面
routes	路由文件，以指定的http请求方式暴露给用户，并在用户请求后将结果返回
index.js	
users.js	
views	视图文件
app.js	初始化文件，引入依赖项
mysql.js	连接mysql

由于express脚手架已经搭建好了前后端的基本框架，我们在此基础上进行路由的修改、功能的实现以及美化即可。需要进行更改的文件有index.js, search.html, time.html, style.css。下面将附上index.js, search.html, style.css的代码，time.html会在时间热度分析时附上。

## index.js

index.js的主要功能是接收从search.html中传入的参数（如标题的参数为request.query.title），然后根据参数进行sql语句的查询，并将查询结果通过response返回给search.html。此处主要实现的是对标题和时间两个字段的查询。需要注意的一个问题是，用户在前端页面查询时可能会空缺某个字段从而产生一些空值，可以通过修改sql查询语句或者通过if条件控制避免出错，并且达到不管哪些项为空值，都能进行查询（全为空则返回所有结果）。此外，在写sql语句时要注意，由于语句较长，往往需要分成几行写，每一行的末尾要记得加上空格，某些字段要记得加上引号，在写完sql语句之后最好先到数据库中尝试执行一下，执行成功之后再放入index.js文件中。

```
//index.js
var express = require('express');
var router = express.Router();
var mysql = require('../mysql.js');

/* GET home page. */
router.get('/', function(req, res, next) {
  res.render('index', { title: 'Express' });
});
//与search.html相关联
router.get('/process_get1', function(request, response) {
  //sql字符串和参数
  if (!request.query.publish_date) {
    var fetchSql = "select url,source_name,title,publish_date " +
      "from eastmoney where title like '%" + request.query.title + "%' " +
      "order by publish_date desc";
  }
  else{
    var fetchSql = "select url,source_name,title,publish_date " +
      "from eastmoney where title like '%" + request.query.title + "%' " +

      "and publish_date like '" + request.query.publish_date + "' " +
      "order by publish_date desc";
  }
  mysql.query(fetchSql, function(err, result, fields) {
    for (var i=0;i<result.length;i++){
      result[i].publish_date = result[i].publish_date.toLocaleDateString()
    }
    response.writeHead(200, {
```

```

        "Content-Type": "application/json"
    });
    response.write(JSON.stringify(result));
    response.end();
});
});
//与time.html相关联
router.get('/process_get2', function(request, response) {
    //sql字符串和参数

    if (!request.query.publish_date1 && !request.query.publish_date2) {
        var fetchSql = "select publish_date ,COUNT(*) AS `num` " +
            "from eastmoney where content like '%" + request.query.content + "%' " +

            "GROUP BY publish_date " +
            "order by publish_date desc";
    }
    else if(!request.query.publish_date1){
        var fetchSql = "select publish_date ,COUNT(*) AS `num` " +
            "from eastmoney where content like '%" + request.query.content + "%'
" +

            "and date(publish_date) <= '" + request.query.publish_date2 + "' " +
            "GROUP BY publish_date " +
            "order by publish_date desc";
    }
    else if(!request.query.publish_date2){
        var fetchSql = "select publish_date ,COUNT(*) AS `num` " +
            "from eastmoney where content like '%" + request.query.content + "%'
" +

            "and date(publish_date) >= '" + request.query.publish_date1 + "' " +
            "GROUP BY publish_date " +
            "order by publish_date desc";
    }
    else{
        var fetchSql = "select publish_date ,COUNT(*) AS `num` " +
            "from eastmoney where content like '%" + request.query.content + "%'
" +

            "and date(publish_date) between '" + request.query.publish_date1 + "'
and '" + request.query.publish_date2 + "' " +
            "GROUP BY publish_date " +
            "order by publish_date desc";
        console.log(fetchSql);
    }
    mysql.query(fetchSql, function(err, result, fields) {
        for (var i=0;i<result.length;i++){
            result[i].publish_date = result[i].publish_date.toLocaleDateString()
        }
        response.writeHead(200, {
            "Content-Type": "application/json"
        });
        response.write(JSON.stringify(result));
        response.end();
    });
});
module.exports = router;

```

## search.html

search.html 需要提供一个表单到前端网页供用户输入查询参数，然后网页会将用户输入的参数进行返回给HTML(此处通过get方法)，HTML解析参数并传递给index.js，最后将index.js返回的查询结果以表格的形式展现到网页中。以及，此处使用了bootstrap进行分页，但此处的代码框架并非我写，而是调用了网上已经写好的css和js文件完成了分页的功能并且以表格的形式进行显示。我写的不分页的代码在github项目代码中进行了注释，也可以很好的进行表格展示。

```
//search.html
<!DOCTYPE html>
<html>
<header>
  <link href="./style.css" rel="stylesheet" type="text/css"/>
  <script src="https://cdn.bootcss.com/jquery/3.4.1/jquery.js"></script>
  <link href="http://www.itxst.com/package/bootstrap-table-1.14.1/bootstrap-4.3.1/css/bootstrap.css" rel="stylesheet" />
  <link href="http://www.itxst.com/package/bootstrap-table-1.14.1/bootstrap-table-1.14.1/bootstrap-table.css" rel="stylesheet" />
  <script src="http://www.itxst.com/package/bootstrap-table-1.14.1/bootstrap-table-1.14.1/bootstrap-table.js"></script>
</header>

<body>
  <div class="notboot">财经新闻查询</div>
  <div>
    <form>
      标题: <input type="text" name="title_text" id="query1"> 时间: <input type="date" name="publish_date" id="query2"> <input class="form-submit" type="button" value="查询">
    </form>
  </div>
  <table width="100%" id="record2"></table>
  <script>
    $(document).ready(function() {
      //点击查询按钮之后会进行以下操作
      $("input:button").click(function() {
        $.get('/process_get1?title=' + $("#query1").val() + '&publish_date=' + $("#query2").val(), function(data) {
          $("#record2").bootstrapTable({
            search:false,          //加上搜索控件
            method: 'get',         //请求方式
            cache: false,          //是否使用缓存，默认为true，所以一般情况下需要设置一下这个属性（*）
            pagination: true,      //是否显示分页（*）
            sortable: true,        //是否启用排序
            sortOrder: "asc",      //排序方式
            striped: true,         //是否显示行间隔色
            uniqueId: "url",       //每一行的唯一标识，一般为主键列
            pageSize: 5,           //每页的记录行数
            sidePagination: 'client',
            columns:[{
              field:'url',
              title:'链接'
            },{
              field:'source_name',
              title:'来源'
            },{

```



```
        field: 'title',
        title: '标题'
    }, {
        field: 'publish_date',
        title: '发表日期'
    }],
    data: data,
  });
});
});
</script>
</body>

</html>
```

### style.css

style.css主要是为了美化html页面，就不放在此处浪费空间了。

### 结果展示

在cmd中输入node bin/www，然后本地访问<http://127.0.0.1:3000/search.html>，即可看到结果。

分页之前

财经新闻查询

标题: 股 时间: 2021/04/28 查询

number	url	source_name	title	publish_date
1	http://finance.eastmoney.com/a/202104281903569355.html	东方财富网	中报业绩预计翻倍增长股名单出炉 最牛股业绩增逾90倍 _ 东方财富网	2021/4/28
2	http://finance.eastmoney.com/a/202104281904729598.html	东方财富网	寿仙谷一季度扣非净利增14% 今股价跌1.6% _ 东方财富网	2021/4/28
3	http://finance.eastmoney.com/a/202104281904727042.html	东方财富网	传音控股首季净利8亿 去年ROE手机产品毛利率均降 _ 东方财富网	2021/4/28
4	http://finance.eastmoney.com/a/202104281903442915.html	东方财富网	港股早知道: 中国铝业一季度归母净利润9.67亿元 同比大增3025.68% _ 东方财富网	2021/4/28
5	http://finance.eastmoney.com/a/202104281904680681.html	东方财富网	巨丰投顾: 贵州茅台放量下挫 谁来打破A股盘整格局? _ 东方财富网	2021/4/28
6	http://finance.eastmoney.com/a/202104281903447626.html	东方财富网	悔青了 高毅资产错失有色龙头! 最“痴情”股长拿三年 这些绩优股被“错卖”? _ 东方财富网	2021/4/28
7	http://finance.eastmoney.com/a/202104281903419098.html	东方财富网	热门中概股涨跌互现 线下教育股领跌 _ 东方财富网	2021/4/28
8	http://finance.eastmoney.com/a/202104281903701048.html	东方财富网	北向资金3日扫货超百亿 加仓这些个股 (名单) _ 东方财富网	2021/4/28
9	http://finance.eastmoney.com/a/202104281903390833.html	东方财富网	欧股主要指数小幅收低 德国DAX30指数跌0.31% _ 东方财富网	2021/4/28
10	http://finance.eastmoney.com/a/202104281903328710.html	东方财富网	云南股王“炒股”亏损近8亿 A股一哥名头再登中纪委网站 发生了什么? _ 东方财富网	2021/4/28
11	http://finance.eastmoney.com/a/202104281904729353.html	东方财富网	ST中安: 控股股东中恒汇志所持约5.28亿股将被冻结 _ 东方财富网	2021/4/28
12	http://finance.eastmoney.com/a/202104281904550279.html	东方财富网	退市新规“发威”! 一日内5家公司因财务不达标被实施“ST 创A股新纪录 _ 东方财富网	2021/4/28
13	http://finance.eastmoney.com/a/202104281904623055.html	东方财富网	惠阳银行4.67亿定增发行实施 前十大股东有调整 _ 东方财富网	2021/4/28

分页之后 (每页5项内容)

## 财经新闻查询

财经新闻查询

标题:

A股

时间:

2021/04/28



查询

链接	来源	标题	发表日期
<a href="http://finance.eastmoney.com/a/202104281904680681.html">http://finance.eastmoney.com/a/202104281904680681.html</a>	东方财富网	巨丰投顾：贵州茅台放量下挫 谁来打破A股盘整格局？_东方财富网	2021/4/28
<a href="http://finance.eastmoney.com/a/202104281903328710.html">http://finance.eastmoney.com/a/202104281903328710.html</a>	东方财富网	云南股王“炒股”亏损近8亿 A股一哥名头再登中纪委网站 发生了什么？_东方财富网	2021/4/28
<a href="http://finance.eastmoney.com/a/202104281904550279.html">http://finance.eastmoney.com/a/202104281904550279.html</a>	东方财富网	退市新规“发威”！一日内5家公司因财务不达标被实施“ST” 创A股新纪录_东方财富网	2021/4/28
<a href="http://finance.eastmoney.com/a/202104281903706781.html">http://finance.eastmoney.com/a/202104281903706781.html</a>	东方财富网	大基金与小米长江基金先后入局 这家代表先进制造业的A股牛气了_东方财富网	2021/4/28
<a href="http://stock.caijing.com.cn/20210428/4760738.shtml">http://stock.caijing.com.cn/20210428/4760738.shtml</a>	财经网	逾四成A股投资者一季度盈利 创近1年来单季盈利者占比最低_市场动态_资本市场_财经网 - CAIJING.COM.CN	2021/4/28

Showing 1 to 5 of 9 rows

<

1

2

>

如果不输入时间，则会将所有天包含该关键词的新闻都显示出来，并且按照发表日期进行排序，越晚发表的新闻越靠前（这一功能的实现会在下面进行说明）。

## 财经新闻查询

财经新闻查询

标题:

A股

时间:

年 / 月 / 日

查询

链接	来源	标题	发表日期
http://finance.eastmoney.com/a/202104291905460423.html	东方财富网	险资拥抱A股核心资产 国寿连续加仓贵州茅台_东方财富网	2021/4/29
http://finance.eastmoney.com/a/202104291906053873.html	东方财富网	金融圈刷屏！尝尽人间烟火 私募老总转型开滴滴：人到中年 只能是务实！网友：A股太难了_东方财富网	2021/4/29
http://finance.eastmoney.com/a/202104281904680681.html	东方财富网	巨丰投顾：贵州茅台放量下挫 谁来打破A股盘整格局？_东方财富网	2021/4/28
http://finance.eastmoney.com/a/202104281903328710.html	东方财富网	云南股王“炒股”亏损近8亿 A股一哥名头再登中纪委网站 发生了什么？_东方财富网	2021/4/28
http://finance.eastmoney.com/a/202104281904550279.html	东方财富网	退市新规“发威”！一日内5家公司因财务不达标被实施“ST” 创A股新纪录_东方财富网	2021/4/28

Showing 6 to 10 of 30 rows

5 rows per page

<

1

2

3

4

5

6

>

## 时间热度分析

本项目的`时间热度分析`是指，可以显示某个关键词在每一天中出现的次数，以此可以看出该词在哪天是热词在哪天无人问津，从而可以大概分析出每天的焦点事件是什么。其实词语热度分析采用图表更为合适，但是由于我只连续的爬取了3天网页，所有网页大都集中在在这三天，其他天数的很少，图表也不能完全展示出其波动效果，所以使用了表格进行展示，如果长期爬取新闻的话，可以以折线图的方式展现词语的时间热度。

此功能需要新增`time.html`文件并修改`index.js`和`css`文件，`index.js`在前文已经进行了展示，就不再此陈列了，但有一点值得注意，由于时间数据在`mysql`数据库中和网页中是按照不同的时区时间的，所以日期可能会不一致，如下图所示，所以在从数据库中取出时间数据的时候，需要使用`.toLocaleDateString()`函数，将时间数据变成字符串，从而使得前后端时间数据一致。



标题: A股

时间: 2021/04/28

查询

url

<http://finance.eastmoney.com/a/202104281904680681.html>  
<http://finance.eastmoney.com/a/202104281903328710.html>  
<http://finance.eastmoney.com/a/202104281904550279.html>  
<http://finance.eastmoney.com/a/202104281903706781.html>  
<http://stock.caijing.com.cn/20210428/4760738.shtml>

source\_name

东方财富网  
 东方财富网  
 东方财富网  
 东方财富网  
 财经网

title

巨丰投顾: 贵州茅台放量下挫 谁来打破A股盘整格局? \_ 东方财富网  
 云南股王“炒股”亏损近8亿 A股一哥名头再登中纪委网站 发生了什么? \_ 东方财富网  
 退市新规“发威”! 一日内5家公司因财务不达标被实施\*ST 创A股新纪录 \_ 东方财富网  
 大基金与小米长江基金先后入局 这家代表先进制造业的A股牛气了 \_ 东方财富网  
 逾四成A股投资者一季度盈利 创近1年来单季盈利者占比最低\_市场动态\_资本市场\_财经网 - CAIJING.COM.CN

publish\_date

2021-04-27T16:00:00.000Z  
 2021-04-27T16:00:00.000Z  
 2021-04-27T16:00:00.000Z  
 2021-04-27T16:00:00.000Z  
 2021-04-27T16:00:00.000Z

C:\Windows\System32\cmd.exe - mysql -u root -p

mysql> select title,publish\_date from eastmoney where title like '%A股%' and publish\_date like '2021-04-28' order by publish\_date desc;

title	publish_date
巨丰投顾: 贵州茅台放量下挫 谁来打破A股盘整格局? _ 东方财富网	2021-04-28
云南股王“炒股”亏损近8亿 A股一哥名头再登中纪委网站 发生了什么? _ 东方财富网	2021-04-28
退市新规“发威”! 一日内5家公司因财务不达标被实施*ST 创A股新纪录 _ 东方财富网	2021-04-28
大基金与小米长江基金先后入局 这家代表先进制造业的A股牛气了 _ 东方财富网	2021-04-28
逾四成A股投资者一季度盈利 创近1年来单季盈利者占比最低_市场动态_资本市场_财经网 - CAIJING.COM.CN	2021-04-28

5 rows in set (0.00 sec)

mysql>

下面主要介绍time.html的功能，该文件同search.html的框架一样，也是将网页的输入参数传递给index.js然后将index.js返回的结果返回到网页中，它需要提供一个表单到前端网页供用户输入参数，然后将结果以表格的形式返回。表格的内容为该关键词在每一天出现的次数。

考虑到时间越靠近现在，信息就越重要，所以表格按照时间顺序进行展示，发表时间越晚（越靠近现在）就越靠前。这一功能在index.js中进行实现，主要是通过sql查询语句中增加"order by publish\_date desc"这一语句进行排序。

## time.html

```

<!DOCTYPE html>
<html>
<header>
  <link href="./style.css" rel="stylesheet" type="text/css"/>
  <script src="https://cdn.bootcss.com/jquery/3.4.1/jquery.js"></script>
</header>

<body>
  <div class="notboot">财经新闻时间热点查询</div>
  <div>
    <form>
      <br> 查询标题: <input type="text" name="title_text" id="query1">
      <br> 开始时间: <input type="date" name="publish_date1" id="query2">
      <br> 结束时间: <input type="date" name="publish_date2" id="query3">
      <br> <input class="form-submit" type="button" value="查询">
    </form>
  </div>
  <div class="cardLayout" style="margin: 10px 0px">
    <table id="record3" class="tabtop13"></table>
  </div>
  <script>
    $(document).ready(function() {
      //点击查询按钮之后会进行以下操作
      $("input:button").click(function() {
        $.get('/process_get2?title=' + $("#query1").val() +
          '&publish_date1=' + $("#query2").val() + '&publish_date2=' + $("#query3").val(),
          function(data) {
            $("#record3").empty();
            $("#record3").append('<tr><th>publish_date</th>
            <th>number</th></tr>');
            for (let list of data) {

```

```
        let table = '<tr>';
        Object.values(list).forEach(element => {
            table += ('<td>' + element + '</td>');
        });
        $("#record3").append(table + '</tr>');
    }
});

});
</script>
</body>

</html>
```

## 结果展示

最开始我是按照标题进行关键词匹配的，但是考虑到标题的字数有限，很多有意义的词语都没有包含在内，所以改为按照文章内容进行关键词匹配，只要文章中出现了该词语，则为该词贡献了热度，这也符合我们的生活常识。由以下两张图可以看出，用内容匹配出的结果要多很多。

在cmd中输入node bin/www，然后本地访问<http://127.0.0.1:3000/time.html>，即可看到结果。

按照标题进行关键词匹配：（表格按照发表日期进行排序）

### 财经新闻时间热点查询

查询内容：

开始时间： 

结束时间： 

### 查询词在每一天出现的次数

publish_date	number
2021/4/29	7
2021/4/28	9
2021/4/27	9
2021/4/1	2
2021/3/15	1
2021/3/8	2

按照文章内容进行关键词匹配：

## 财经新闻时间热点查询

查询内容:

开始时间:  

结束时间:  

### 查询词在每一天出现的次数

publish_date	number
2021/4/29	35
2021/4/28	52
2021/4/27	43
2021/4/26	2
2021/4/13	1
2021/3/23	1

输入标题和开始结束时间后:

## 财经新闻时间热点查询

查询内容:

开始时间:  

结束时间:  

### 查询词在每一天出现的次数

publish_date	number
2021/4/29	35
2021/4/28	52
2021/4/27	43

## 总结

通过本次实验, 我知道了如何通过node.js进行爬虫, 并在研究html页面时对标签有了进一步的了解, 也学会使用正则表达式匹配合适的文字与链接。此外我还了解到网站前后端的框架是怎样的, 用户在网页中输入的参数是如何被HTML进行解析并传递到node.js中的, 然后又怎样通过node.js与mysql数据库进行连接并对数据进行存储或查询的, 以及从数据库中得到的数据是怎样传回到HTML并在页面以特定的格式进行显示的。此外, 我还学会使用css文件对html文件进行美化, 对前端有了更进一步的了解。

