Jsoup爬虫

一、介绍

jsoup 是一款 Java 的 HTML 解析器，可直接解析某个 URL 地址、HTML 文本内容。它提供了一套非常省力的 API，可通过 DOM、CSS 以及类似取出和操作数据。

1.从一个URL，文件或字符串中解析HTML；

2.使用DOM或CSS选择器来查找、取出数据；

3.可操作HTML元素、属性、文本；

htmlunit是一款开源的Java页面分析工具，读取页面后，可以有效的使用htmlunit 分析页面上的内容。项目可以模拟浏览器运行，被誉为Java浏览界面的浏览器，运行速度也是非常迅速的。

二、依赖引用

```
<!-- jsoup -->
<dependency>
    <groupId>org.jsoup</groupId>
    <artifactId>jsoup</artifactId>
    <version>1.11.3</version>
</dependency>
<!-- htmlunit -->
<dependency>
    <groupId>net.sourceforge.htmlunit</groupId>
    <artifactId>htmlunit</artifactId>
    <version>2.34.1</version>
</dependency>
```

三、爬取图片

1.下载工具类

```
public class JsoupUtils {
    public static void downImages(String filePath, String fileName, String imgUrl) {
        // 若指定文件夹没有，则先创建
        File dir = new File(filePath);
        if (!dir.exists()) {
            dir.mkdirs();
        }
        // 写出的路径
        File file = new File(filePath + File.separator + fileName);
        try {
            // 获取图片URL
            URL url = new URL(imgUrl);
            // 获得连接
            URLConnection connection = url.openConnection();
            // 设置10秒的相应时间
            connection.setConnectTimeout(10 * 1000);
            // 获得输入流
            InputStream in = connection.getInputStream();
            // 获得输出流
            BufferedOutputStream out = new BufferedOutputStream(new FileOutputStream(file));
            // 构建缓冲区
            byte[] buf = new byte[1024];
            int size;
            // 写入到文件
            while (-1 != (size = in.read(buf))) {
                out.write(buf, 0, size);
            }
            out.close();
            in.close();
        } catch (MalformedURLException e) {
```

```
            e.printStackTrace();
        } catch (IOException e) {
            e.printStackTrace();
        }
    }

    public static String formatNumber(Integer number){
        return String.format("%s_%05d.jpg", "bd_img", number);
    }

    /**
     * 获取页面文档字串(等待异步JS执行)
     */
    public static Document getDocumentByHtmlUnit(String url) throws Exception {
        WebClient webClient = new WebClient(BrowserVersion.CHROME);
        WebClientOptions options = webClient.getOptions();
        options.setUseInsecureSSL(true);                                        //是否使用不安全的SSL
        options.setThrowExceptionOnScriptError(false);                          //当JS执行出错的时候是否抛出异常
        options.setThrowExceptionOnFailingStatusCode(false);                    //当HTTP的状态非200时是否抛出异常
        options.setActiveXNative(false);                                        //是否允许使用ActiveX
        options.setCssEnabled(false);                                           //是否启用CSS
        options.setJavaScriptEnabled(true);                                     //很重要，启用JS
        options.setTimeout(30000);                                              //设置"浏览器"的请求超时时间
        options.setDoNotTrackEnabled(false);                                    //不跟踪抓取
        webClient.setAjaxController(new NicelyResynchronizingAjaxController()); //很重要，设置支持AJAX
        webClient.setJavaScriptTimeout(30*1000);                                //设置JS执行的超时时间

        HtmlPage htmlPage = webClient.getPage(url);                             //模拟浏览器打开一个目标网址
        webClient.waitForBackgroundJavaScript(50);                              //该方法阻塞线程
        String xml = htmlPage.asXml();
        webClient.close();
        return Jsoup.parse(xml);
    }
}
```

2.测试

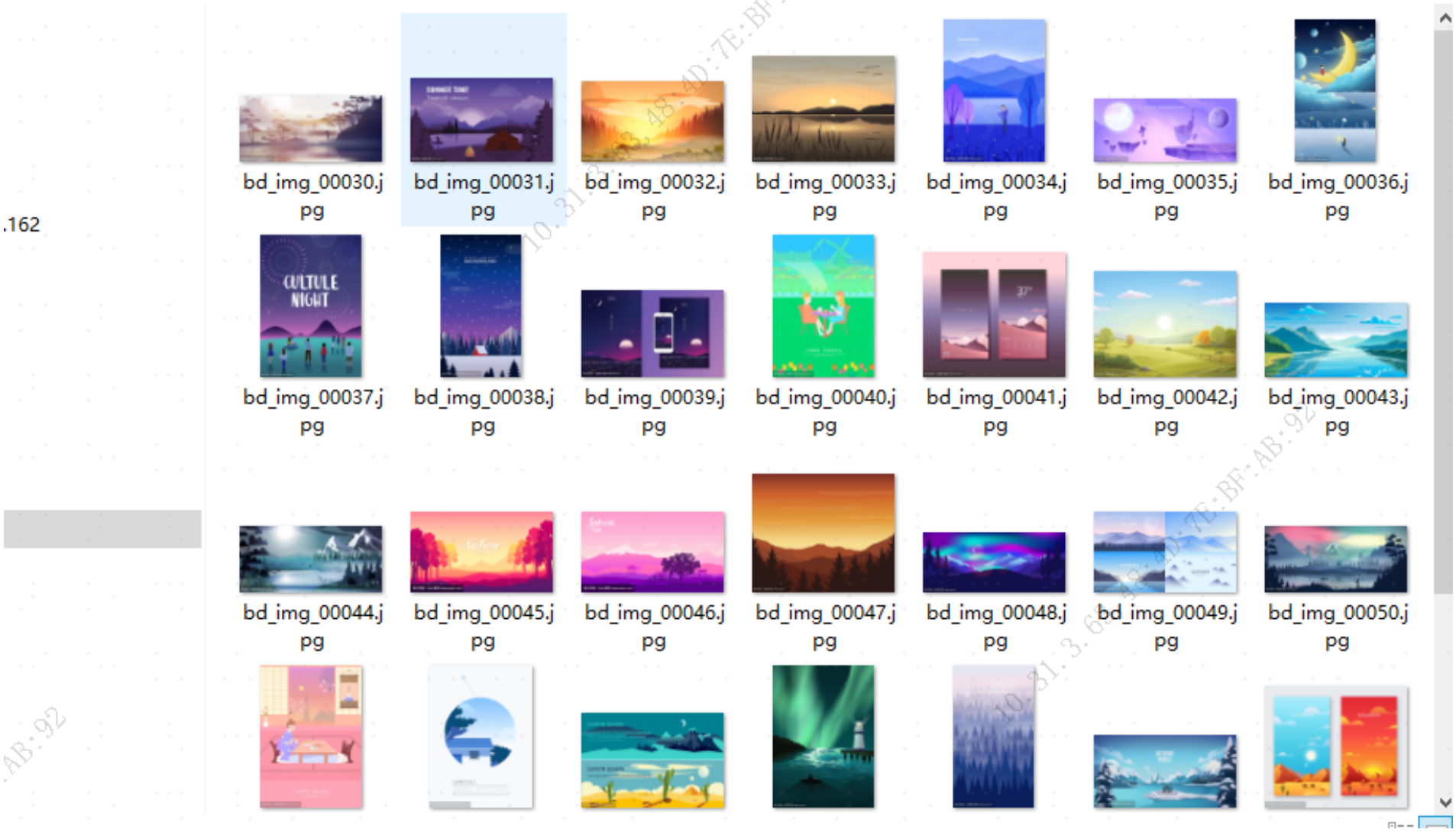
```
@Slf4j
public class JsoupTests extends BaseTest {

    private final static String baseUri = "https://image.baidu.com/";
    private final static String rootDir = "d:/img/";
    private final static Map<String, Map<String, String>> groupMap = new HashMap<>();   //[分组名, [图片名, 图片url]]
    private final static AtomicInteger imgNo = new AtomicInteger(0);

    @Test
    public void testDownImages() throws Exception{
        log.info("开始解析网页...");
        Connection connect = Jsoup.connect(baseUri);
        try {
            //得到Document对象
            Document document = connect.get();

            //图片分组
            Elements groups = document.getElementsByClass("bd-home-content-album-item");
            for (Element group : groups) {
                Elements titles = group.getElementsByClass("bd-home-content-album-item-title");
                String groupTitle = titles.first().text();
                //获取分组详情页面
                String childUri = group.attr("abs:href");

                //Document childDoc = Jsoup.connect(childUri).get();  //获取静态页面
```

```
            Document childDoc = JsoupUtils.getDocumentByHtmlUnit(childUri); //获取动态加载后的内容
            Element imgDiv = childDoc.getElementById("imgList");
            Elements imgList = imgDiv.getElementsByClass("albumsdetail-item-img");
            Map<String, String> imgMap = new HashMap<>();
            for (Element imgElement : imgList) {
                String imgUrl = imgElement.attr("abs:src");
                String imgName = JsoupUtils.formatNumber(imgNo.getAndIncrement());
                imgMap.put(imgName, imgUrl);
            }
            groupMap.put(groupTitle, imgMap);
        }

        //下载图片
        log.info("解析完成，开始下载>>>");
        for (Map.Entry<String, Map<String, String>> stringMapEntry : groupMap.entrySet()) {
            String groupDir = stringMapEntry.getKey();
            if (groupDir.contains(">")) continue;
            log.info("分组下载: {}", groupDir);
            stringMapEntry.getValue().forEach((fileName, fileUrl) ->
                    JsoupUtils.downImages(rootDir.concat(groupDir), fileName, fileUrl)
            );
        }
        log.info("下载完成");
    } catch (IOException e) {
        e.printStackTrace();
    }
  }
}
```

结果

```
Tests passed: 1 of 1 test – 13 sec 561 ms

2022-05-10 14:52:50.776 ERROR 17556 --- [           main] c.g.h.javascript.StrictErrorReporter     : runtimeError: message=[An invalid or illegal selector was
2022-05-10 14:52:50.854  INFO 17556 --- [           main] .g.h.NicelyResynchronizingAjaxController : Re-synchronized call to https://image.baidu.com/user/login
2022-05-10 14:52:50.901  WARN 17556 --- [           main] c.g.htmlunit.IncorrectnessListenerImpl   : Obsolete content type encountered: 'text/javascript'.
2022-05-10 14:52:51.089  WARN 17556 --- [Client@21f7c0be] c.g.htmlunit.IncorrectnessListenerImpl   : Obsolete content type encountered: 'application/x-javascri
2022-05-10 14:52:51.095  INFO 17556 --- [           main] com.formssi.mall.redis.JsoupTests        : 解析完成，开始下载>>>
2022-05-10 14:52:51.095  INFO 17556 --- [           main] com.formssi.mall.redis.JsoupTests        : 分组下载: 渐变风格插画
2022-05-10 14:52:51.625  INFO 17556 --- [           main] com.formssi.mall.redis.JsoupTests        : 分组下载: 城市建筑摄影专题
2022-05-10 14:52:52.231  INFO 17556 --- [           main] com.formssi.mall.redis.JsoupTests        : 分组下载: 航拍地球系列
2022-05-10 14:52:52.703  INFO 17556 --- [           main] com.formssi.mall.redis.JsoupTests        : 分组下载: 宠物图片
2022-05-10 14:52:53.075  INFO 17556 --- [           main] com.formssi.mall.redis.JsoupTests        : 分组下载: 皮影
2022-05-10 14:52:53.751  INFO 17556 --- [           main] com.formssi.mall.redis.JsoupTests        : 下载完成
```

效果


此电脑 › 软件 (D:) › img › 渐变风格插画
在 渐变风格插画 中搜索
bd_img_00030.jpg
bd_img_00031.jpg
bd_img_00032.jpg
bd_img_00033.jpg
bd_img_00034.jpg
bd_img_00035.jpg
bd_img_00036.jpg
bd_img_00037.jpg
bd_img_00038.jpg
bd_img_00039.jpg
bd_img_00040.jpg
bd_img_00041.jpg
bd_img_00042.jpg
bd_img_00043.jpg
bd_img_00044.jpg
bd_img_00045.jpg
bd_img_00046.jpg
bd_img_00047.jpg
bd_img_00048.jpg
bd_img_00049.jpg
bd_img_00050.jpg

页面分析过程：

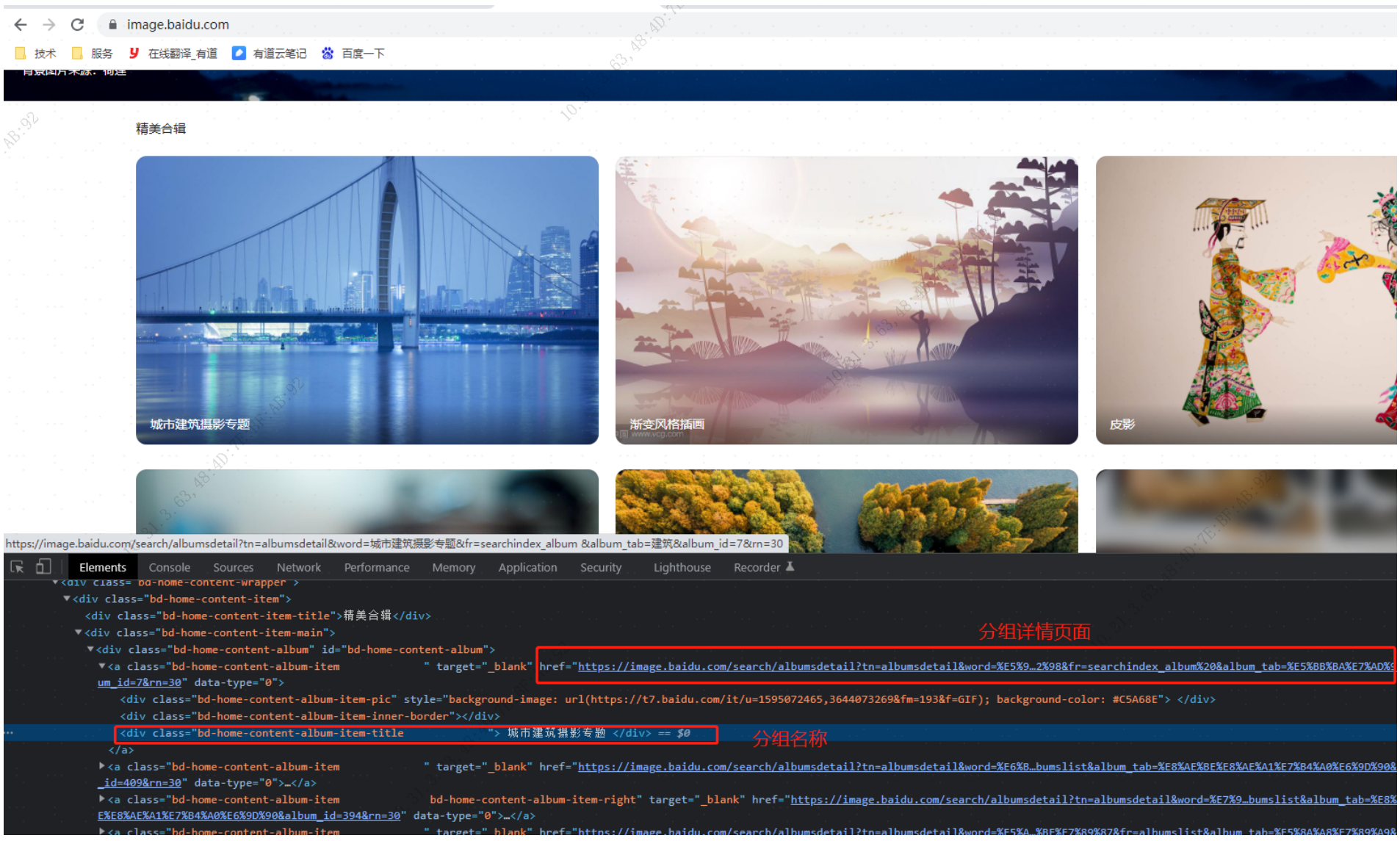


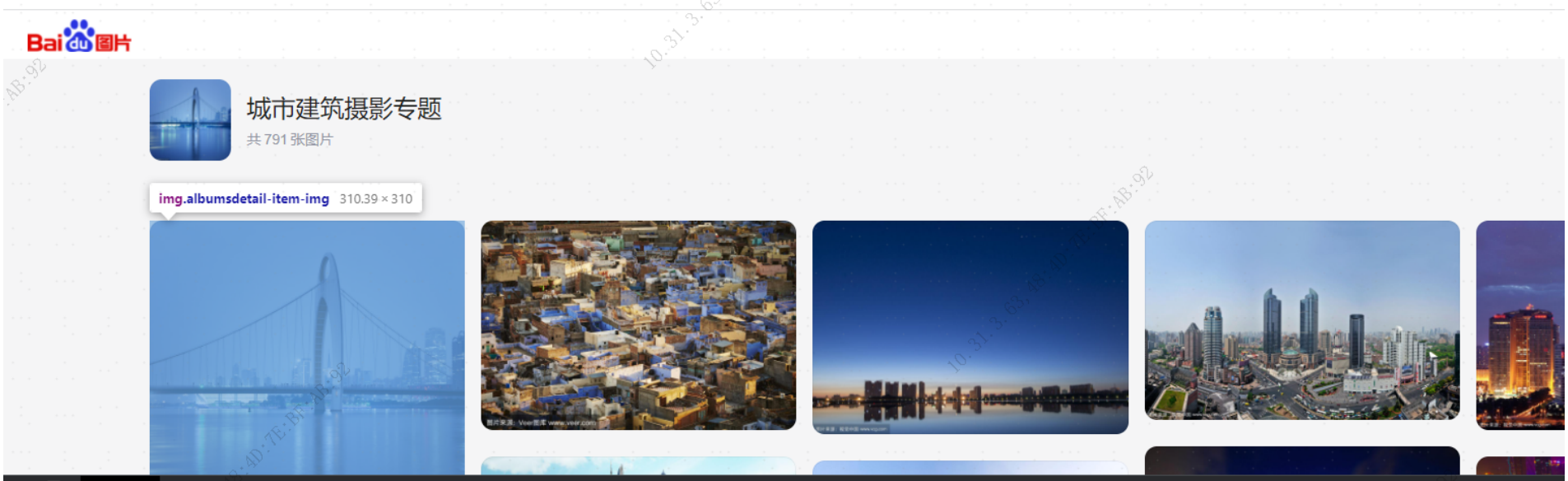
image.baidu.com
技术
服务
在线翻译_有道
有道云笔记
百度一下
精美合辑
城市建筑摄影专题
渐变风格插画
皮影
https://image.baidu.com/search/albumsdetail?tn=albumsdetail&word=城市建筑摄影专题&fr=searchindex_album &album_tab=建筑&album_id=7&rn=30
Elements
Console
Sources
Network
Performance
Memory
Application
Security
Lighthouse
Recorder

精美合辑

<a class="bd-home-content-album-item " target="_blank" href="https://image.baidu.com/search/albumsdetail?tn=albumsdetail&word=%E5%9…2%98&fr=searchindex_album%20&album_tab=%E5%BB%BA%E7%AD%
um_id=7&rn=30" data-type="0">

 城市建筑摄影专题 == $0

<a class="bd-home-content-album-item " target="_blank" href="https://image.baidu.com/search/albumsdetail?tn=albumsdetail&word=%E6%B…bumslist&album_tab=%E8%AE%BE%E8%AE%A1%E7%B4%A0%E6%9D%90&
_id=409&rn=30" data-type="0">…
<a class="bd-home-content-album-item bd-home-content-album-item-right" target="_blank" href="https://image.baidu.com/search/albumsdetail?tn=albumsdetail&word=%E7%9…bumslist&album_tab=%E8%
E%E8%AE%A1%E7%B4%A0%E6%9D%90&album_id=394&rn=30" data-type="0">…
分组详情页面
分组名称


Baidu图片
城市建筑摄影专题
共 791 张图片
img.albumsdetail-item-img 310.39 × 310
Elements
Console
Sources
Network
Performance
Memory
Application
Security
Lighthouse
Recorder
<script type="text/javascript">…</script>

…

…
…
…

此为动态加载标签

<a class="albumsdetail-item" href="/search/detail?tn=baiduimagedetail&word=%E5%9F%8E%E5%B8%82%E5%BB%BA%E7%AD…%2F%2Ft7.baidu.com%2Fit%2Fu%3D1595072465%2C3644073269%26fm%3D193%26f%3DGIF" target="_blank
index="0" width="310.4" style="width: 310.4px; height: 310px;">
<img class="albumsdetail-item-img" src="https://t7.baidu.com/it/u=1595072465,3644073269&fm=193&f=GIF" style="width: 310.4px; height: 310px; background-color: rgb(197, 166, 142);"> == $0

<a class="albumsdetail-item" href="/search/detail?tn=baiduimagedetail&word=%E5%9F%8E%E5%B8%82%E5%BB%BA%E7%AD…%2F%2Ft7.baidu.com%2Fit%2Fu%3D3779234486%2C1094031034%26fm%3D193%26f%3DGIF" target="_blank
index="9" width="310.4" style="width: 310.4px; height: 206px;">…
<a class="albumsdetail-item" href="/search/detail?tn=baiduimagedetail&word=%E5%9F%8E%E5%B8%82%E5%BB%BA%E7%AD…A%2F%2Ft7.baidu.com%2Fit%2Fu%3D318887420%2C2894941323%26fm%3D193%26f%3DGIF" target="_blank
index="14" width="310.4" style="width: 310.4px; height: 206px;">…
<a class="albumsdetail-item" href="/search/detail?tn=baiduimagedetail&word=%E5%9F%8E%E5%B8%82%E5%BB%BA%E7%AD…%2F%2Ft7.baidu.com%2Fit%2Fu%3D1285847167%2C3193778276%26fm%3D193%26f%3DGIF" target="_blank
index="19" width="310.4" style="width: 310.4px; height: 206px;">…
<a class="albumsdetail-item" href="/search/detail?tn=baiduimagedetail&word=%E5%9F%8E%E5%B8%82%E5%BB%BA%E7%AD…A%2F%2Ft7.baidu.com%2Fit%2Fu%3D3124693600%2C356058981%26fm%3D193%26f%3DGIF" target="_blank