# Ali 1688产品数据抓取分析

## 前置分析

### re.1688.com 分析：

#### 列表页面：

列表数据：https://data.p4psearch.1688.com/data/ajax/get\_premium\_offer\_list.json?beginpage=1

#### 详情页面：

var iDetailConfig

var iDetailData

iDetailData.registerRenderData（类目相关信息）

https://tui.taobao.com/recommend?callback=jQuery (nothing)

#### 描述信息：

https://img.alicdn.com/tfscom/TB1AvjAy1H2gK0jSZJnXXaT1FXa

https://desc.alicdn.com/i3/610/661/611668832029/TB1o\_YXvrH1gK0jSZFw8qw7aXla.desc%7Cvar%5Edesc%3Blang%5Egbk%3Bsign%5E8596346ba80d303f23c505d92243ddac%3Bt%5E1582868241

https://laputa.1688.com/offer/ajax/widgetList.do?callback= (运费)

### kj.1688.com 分析：

#### 列表页面：

列表数据:

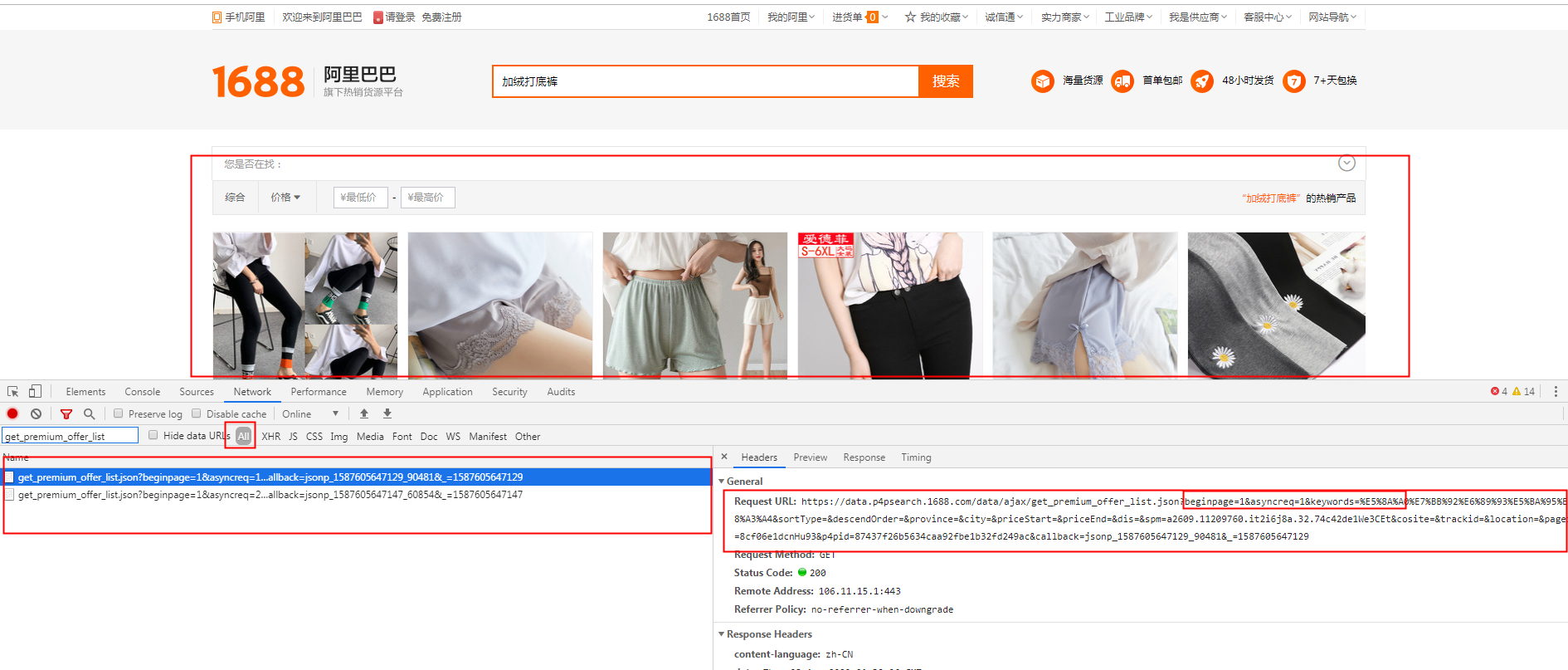
https://s.1688.com/selloffer/rpc\_async\_render.jsonp?keywords=%D0%DD%CF%D0%BF%E3+%C5%AE&n=y&templateConfigName=overseasOfferresult&offset=0&pageSize=60&asyncCount=20&startIndex=0&async=true&enableAsync=true&rpcflag=new&\_pageName\_=%BF%E7%BE%B3%CB%D1%CB%F7&requestId=&callback=jQuery17208870998838762245\_1588747366557

https://s.1688.com/selloffer/rpc\_async\_render.jsonp?keywords=%D0%DD%CF%D0%BF%E3+%C5%AE&n=y&templateConfigName=overseasOfferresult&offset=0&pageSize=60&asyncCount=20&startIndex=20&async=true&enableAsync=true&rpcflag=new&\_pageName\_=%BF%E7%BE%B3%CB%D1%CB%F7&requestId=&callback=jQuery17208870998838762245\_1588747366552

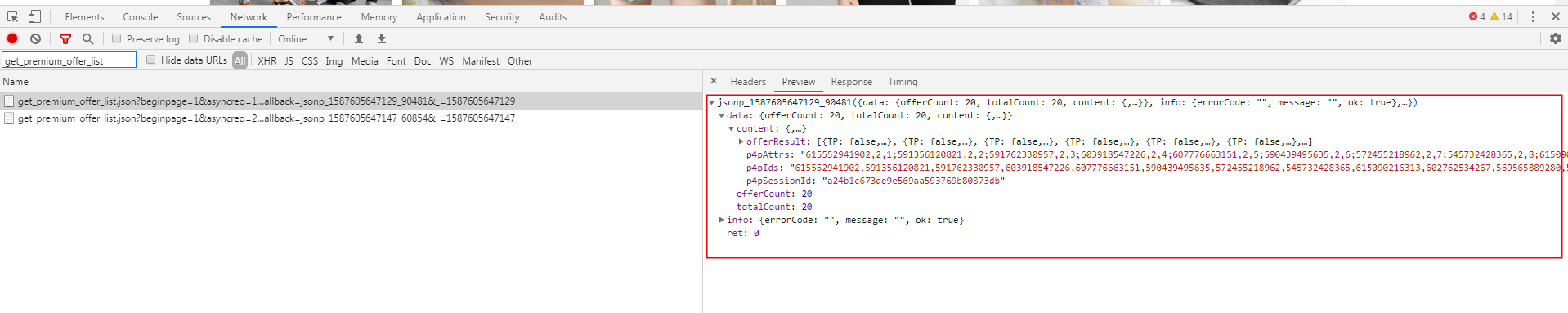
https://s.1688.com/selloffer/rpc\_async\_render.jsonp?keywords=%D0%DD%CF%D0%BF%E3+%C5%AE&n=y&templateConfigName=overseasOfferresult&offset=0&pageSize=60&asyncCount=20&startIndex=40&async=true&enableAsync=true&rpcflag=new&\_pageName\_=%BF%E7%BE%B3%CB%D1%CB%F7&requestId=&callback=jQuery17208870998838762245\_1588747366557

## step 1 分析产品列表

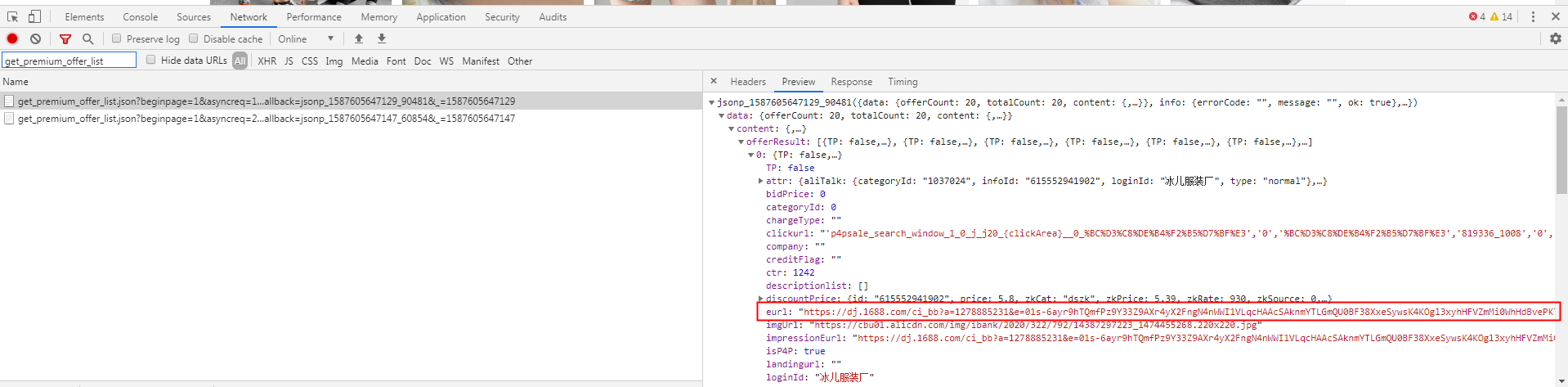
首先需要进入到产品列表页面，如： <https://p4psearch.1688.com/p4p114/p4psearch/offer.htm?spm=a2609.11209760.it2i6j8a.32.74c42de1We3CEt&cosite=&keywords=%E5%8A%A0%E7%BB%92%E6%89%93%E5%BA%95%E8%A3%A4&trackid=&location>= ；根据产品列表页面的每页的产品数据来获取每个产品的明细信息。1688的产品列表的产品数据并不是和页面同步返回的，而是在页面渲染完成后通过Ajax(使用jsonp)异步请求回来渲染的。经分析发现，此Ajax请求的URL如：<https://data.p4psearch.1688.com/data/ajax/get_premium_offer_list.json?beginpage=1&asyncreq=1&keywords=%E5%8A%A0%E7%BB%92%E6%89%93%E5%BA%95%E8%A3%A4&sortType=&descendOrder=&province=&city=&priceStart=&priceEnd=&dis=&spm=a2609.11209760.it2i6j8a.32.74c42de1We3CEt&cosite=&trackid=&location=&pageid=8cf06e1dcnHu93&p4pid=87437f26b5634caa92fbe1b32fd249ac&callback=jsonp_1587605647129_90481&_=1587605647129> ；其中参数beginpage为页码，即第几页；syncreq是每一页的一部分，值都是1到 6，即每一页的数据又分了6次请求，其中其他参数不变，只有syncreq的值在变，每一次请求都是返回20条数据，即每页最多有20 \* 6 = 120条数据；keywords是产品列表数据的显示的关键词。在返回的数据中有每一个产品的明细地址，如下图：



每一页请求数据图



返回的数据图



数据产品明细URL图

## step 2 模拟构造产品列表请求

经step 1分析，产品数据时不能直接从页面DOM节点中获取到的，需要模拟发送Ajax请求才能获取到数据。产品列表数据的请求URL都是类似于 <https://data.p4psearch.1688.com/data/ajax/get_premium_offer_list.json?beginpage=1&asyncreq=1&keywords>=连衣裙 ， 这样的地址，只需要把beginpage,asyncreq,keywords的值改变即可，如抓取前10页的数据，即：beginpage为1到10，而每一页又分asyncreq小片数据，asyncreq的值为 1 到 6，所有10页的数据需要请求10 \* 6 = 60 次才能完成。从返回的数据来看，其中data.content. offerResult 即是产品列表数据，其中每一个对象中的eurl属性即是产品的明细页面地址。

## step 3 构造产品明细地址请求

经过step 2后，可收集到所有的产品明细地址，如：

<https://dj.1688.com/ci_bb?a=1278885231&e=01s-6ayr9hTQmfPz9Y33Z9AXr4yX2FngN4nWWI1VLqcHAAcSA>...

将所有的明细地址发送请求，可以返回产品明细页面，然后对页面的HTML代码进行分析解析数据。

## step 4 分析解析产品明细数据

对获取到HTML页面代码后进行分析，可将产品分为：产品基础信息、卖家信息、SKU信息、明细信息（包含 产品属性、产品描述）、相册信息。

### 基础产品信息

#### 基础信息



如上图所示，基础产品信息如标题、价格等都可以通过解析HTML获取到

#### 类目信息

经发现有段JS代码包含了类目信息，可根据关键词截取到有用的片段并解析成JSON数据即可提取出类目信息，如下图：



### 卖家信息

发现有段JS代码包含了卖家信息，可根据关键词截取到有用片段并解析成JSON数据即可，其中包含了卖家部分信息、产品部分信息,如下图：



也可以通过如下图所示区域，可解析卖家的部分信息



### SKU信息

经发现有片段JS代码中包含了SKU的相关信息，根据关键词截取到有用的片段并解析成JSON数据，如下图：



### 明细信息

这里的明细包含了产品属性、产品描述两个部分

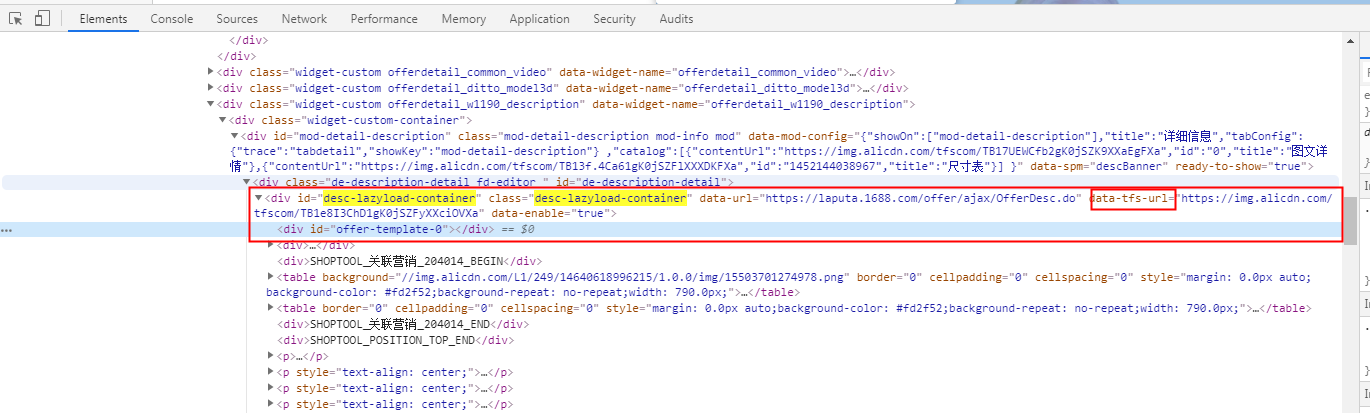
#### 产品属性

产品属性可以通过解析HTML代码提取出来，具体看解析代码，如下图：



#### 产品描述

经分析发现产品描述数据也是通过Ajax请求来的，发现请求明细的Ajax地址是元素的属性值，如下图：



其中，id=” desc-lazyload-container” 元素的data-tfs-url 属性的值就是获取描述信息的地址，所以模拟请求描述信息，即可获取到描述信息。

### 相册信息

相册信息可以通过解析页面相关区域的信息即可获得，如下图：

