数据抓取案例：

抓取网页<https://item.jd.com/7418428.html> 如下数据

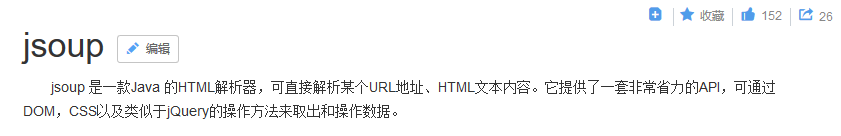


思路：

1. 获得网页源码
2. 网页源代码字符串，解析字符串，然后就开始截取

采用技术：Jsoup

中文网站：<http://www.open-open.com/jsoup/>

实现步骤：

1. 引入依赖或jar包

|  |
| --- |
| <**dependencies**>  <**dependency**>  *<!-- jsoup HTML parser library @ https://jsoup.org/ -->* <**groupId**>org.jsoup</**groupId**>  <**artifactId**>jsoup</**artifactId**>  <**version**>1.11.3</**version**>  </**dependency**> </**dependencies**> |

1. 代码实现

|  |
| --- |
| **public class** UrlSpide {   **public static void** main(String[] args) **throws** Exception{  *//https://search.jd.com/Search?keyword=%E5%8D%8E%E7%A1%95%E7%94%B5%E8%84%91&enc=utf-8&wq=%E5%8D%8E%E4%B8%BA%E7%94%B5%E8%84%91&pvid=ff2f9b2d524e458ebf2a22de3032e6a1* List<String> urlList = **new** ArrayList<String>();  *//提取网页url* String keyword =**"华硕电脑"**;  String url = **"https://search.jd.com/Search?enc=utf-8&wq=%E5%8D%8E%E4%B8%BA%E7%94%B5%E8%84%91&pvid=675a647cd14e4baaa15a5c11b4643fe9&keyword="**+keyword;  Document doc = Jsoup.*connect*(url).get();  Elements eles = doc.select(**"#J\_goodsList ul li"**);  **for** (**int** i = 0; i < eles.size(); i++) {  Element liElement = eles.get(i);  Element aElement = liElement.selectFirst(**".p-img a"**);  urlList.add(**"http:"**+aElement.attr(**"href"**));  }  **for** (String url\_ : urlList) {  *htmParse*(url\_);  }  }   **public static void** htmParse(String url) **throws** Exception{   List<Map<String,String>> list = **new** ArrayList<Map<String,String>>();*//key - 名称 value 就是值  //1. 获得网页的document对象* Document doc = Jsoup.*connect*(url).get();  *// String strHtml = doc.html();  // System.out.println(strHtml);  //2. 解析网页  // Elements eles = doc.select("#parameter-brand");  // Element element = eles.get(0);  // System.out.println(element.text());* Elements eles = doc.select(**"ul.parameter2.p-parameter-list li"**);  **for** (**int** i = 0; i <eles.size() ; i++) {  Element element = eles.get(i);  String context = element.html();  String[] str = context.split(**"："**);  Map<String,String> map = **new** HashMap<String, String>();  map.put(str[0],str[1]);  list.add(map);  }  System.***out***.println(list);  } } |