# 网络爬虫第一天

## 第一章 概念

### 1.1 是什么

网络爬虫 （web crawler），是按照一定的规则，自动地抓取万维网的信息的程序或脚本。

在大数据时代，信息的采集是一项重要的工作，而互联网中的数据是海量的，如果单纯靠人力进行信息采集，不仅低效繁琐，搜集的成本也会提高。如何自动高效地获取互联网中我们感兴趣的信息并为我们所用是一个重要的问题，而爬虫技术就是为了解决这些问题而生的。

网络爬虫(Web crawler) 也叫做网络机器人，可以代替人们自动地在互联网中进行数据信息的采集与整理。它是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本，可以自动采集所有其能够访问到的页面内容，以获取相关数据。

从功能上来讲，爬虫一般分为数据采集，处理，储存三个部分。爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面.上抽取新的URL放入队列,直到满足系统的一定停止条件。

### 1.2 为什么有web crawler

我们初步认识了网络爬虫，但是为什么要学习网络爬虫呢?只有清晰地知道我们的学习目的，才能够更好地学习这一项知识。在此，总结了4种常见的学习爬虫的原因:。

1.可以实现搜索引擎。

我们学会了爬虫编写之后，就可以利用爬虫自动地采集互联网中的信息，采集回来后进行相应的存储或处理，在需要检索某些信息的时候，只需在采集回来的信息中进行检索，即实现了私人的搜索引擎。

2.大数据时代，可以让我们获取更多的数据源。。

在进行大数据分析或者进行数据挖掘的时候，需要有数据源进行分析。我们可以从某些提供数据统计的网站获得，也可以从某些文献或内部资料中获得，但是这些获得数据的方式，有时很难满足我们对数据的需求，而手动从互联网中去寻找这些数据，则耗费的精力过大。此时就可以利用爬虫技术，自动地从互联网中获取我们]感兴趣的数据内容，并将这些数据内容爬取回来，作为我们的数据源，再进行更深层次的数据分析，并获得更多有价值的信息。

3.可以更 好地进行搜索引擎优化(SEO)。。

对于很多SEO从业者来说，为了更好的完成工作，那么就必须要对搜索引擎的工作原理非常清楚，同时也需要掌握搜索引擎爬虫的工作原理。

而学习爬虫，可以更深层次地理解搜索引擎爬虫的工作原理，这样在进行搜索引擎优化时，才能知己知彼，百战不殆。

4.有利于就业。。

从就业来说，爬虫工程师方向是不错的选择之一，因为目前爬虫工程师的需求越来越大，而能够胜任这方面岗位的人员较少，所以属于一个比较紧缺的职业方向，并且随着大数据时代和人工智能的来临，爬虫技术的应用将越来越广泛，在未来会拥有很好的发展空间。

## 第二章 入门程序

网络爬虫 （web crawler），是按照一定的规则，自动地抓取万维网的信息的程序或脚本。

### 2.1 环境准备

JDK1.8

IDEA

## 第三章 HttpClient

网络爬虫就是用程序帮助我们访问网络上的资源，我们一直以来都是使用HTTP协议访问互联网的网页，网络爬虫需要编写程序，在这里使用同样的HTTP协议访问网页。

这里我们使用Java 的HTTP协议客户端HttpClient 这个技术，来实现抓取网页数据。

### 3.1 无参Get请求

往不存在的topic发数据也可以发，因为会默认创建一个partition和1个replication-factor

### 3.2 带参Get请求

index和log最重要

### 3.3 Post请求

### 3.4 带参数的Post请求

### 3.5连接池

如果每次请求都要创建HttpClient， 会有频繁创建和销毁的问题，可以使用连接池来解决这个问题。

测试以下代码，并断点查看每次获取的HttpClient都是不一-样的。。

public class HttpClientPooltest {  
  
  
 public static void main(String[] args) {  
 *//1.创建连接池管理器* PoolingHttpClientConnectionManager cm = new PoolingHttpClientConnectionManager();  
 *//最大连接数* cm.setMaxTotal(100);  
 *//每个主机的连接数（host，就是你访问的主机。目的：爬取数据不能总爬同一个host）* cm.setDefaultMaxPerRoute(10);  
  
  
 *//使用连接池管理发起请求  
 doGet*(cm);  
 }  
  
 private static void doGet(PoolingHttpClientConnectionManager cm) {  
 CloseableHttpClient httpClient = HttpClients.*custom*().setConnectionManager(cm).build();  
  
  
 HttpGet httpGet = new HttpGet("http://www.itcast.cn");  
  
 CloseableHttpResponse response = null;  
 try {  
 response = httpClient.execute(httpGet);  
  
 if (response.getStatusLine().getStatusCode() == 200) {  
 String content = EntityUtils.*toString*(response.getEntity(), "utf8");  
 System.*out*.println(content.length());  
 }  
 } catch (IOException e) {  
 e.printStackTrace();  
 }finally {  
 try {  
 response.close();  
 *//不能关闭httpClient，由连接池管理httpClient  
// httpClient.close();* } catch (IOException e) {  
 e.printStackTrace();  
 }  
 }  
 }  
  
  
}

### 3.6 请求参数

有时候因为网络，或者目标服务器的原因，请求需要更长的时间才能完成，

我们需要自定义相关时间。

HttpGet httpGet = new HttpGet("http://www.itcast.cn");  
  
*//配置请求信息*RequestConfig config = RequestConfig.*custom*().setConnectTimeout(1000)*//创建链接的最长时间，单位毫秒* .setConnectionRequestTimeout(500)*//设置获取链接的最长时间，单位毫秒* .setSocketTimeout(10\*1000)*//设置数据传输的最长时间，单位毫秒* .build();  
  
httpGet.setConfig(config);  
  
CloseableHttpResponse response = null;

## 第四章 Jsoup

我们抓取到页面之后，还需要对页面进行解析。可以使用字符串处理工具解析页面，也可以使用正则表达式，但是这些方法都会带来很大的开发成本，所以我们需要使用一款专门解析html页面的技术。

### 4.1 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 以及类似于jQuery 的操作方法来取出和操作数据。

jsoup的主要功能如下:。

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元索、属性、文本;。

### 4.2 解析

<dependencies>  
  
 <!--httpclient-->  
 <dependency>  
 <groupId>org.apache.httpcomponents</groupId>  
 <artifactId>httpclient</artifactId>  
 <version>4.5.2</version>  
 </dependency>  
  
  
 <!--jsoup-->  
 <dependency>  
 <groupId>org.jsoup</groupId>  
 <artifactId>jsoup</artifactId>  
 <version>1.10.2</version>  
 </dependency>  
  
 <!--io-->  
 <dependency>  
 <groupId>commons-io</groupId>  
 <artifactId>commons-io</artifactId>  
 <version>2.6</version>  
 </dependency>  
  
 <!--commons-lang3-->  
 <dependency>  
 <groupId>org.apache.commons</groupId>  
 <artifactId>commons-lang3</artifactId>  
 <version>3.7</version>  
 </dependency>  
  
  
  
  
  
 <dependency>  
 <groupId>org.slf4j</groupId>  
 <artifactId>slf4j-log4j12</artifactId>  
 <version>1.7.25</version>  
 <!--<scope>test</scope>-->  
 </dependency>  
  
 <!--junit-->  
 <dependency>  
 <groupId>junit</groupId>  
 <artifactId>junit</artifactId>  
 <version>4.12</version>  
 </dependency>  
</dependencies>

#### 4.2.1 解析url

@Test  
public void testUrl() throws Exception{  
 *//解析url  
 //第一参数：访问的url,第二个参数是访问时的超时时间* Document doc = Jsoup.*parse*(new URL("http://www.itcast.cn"), 1000);  
  
  
 *//使用标签选择器，获取title中的内容* String title = doc.getElementsByTag("title").first().text();  
 System.*out*.println(title);  
}

PS:虽然使用Jsoup可以昔代HttpClient直接发起请求解析数据，但是往往不会这样用，因为实际的开发过程中，需要使用到多线程，连接池，代理等等方式，而jsoup对这些的支持并不是很好，所以我们一般把jsoup 仅仅作为Html解析工具使用。

#### 4.2.2 解析字符串

先准备以下html文件

|  |
| --- |
| <html>  <head>  <title>传智播客官网-一样的教育,不一样的品质</title>  </head>  <body>  <div class="city">  <h3 id="city\_bj">北京中心</h3>  <fb:img src="/2018czgw/images/slogan.jpg" class="slogan"/>  <div class="city\_in">  <div class="city\_con" style="display: none;">  <ul>  <li id="test" class="class\_a class\_b">  <a href="http://www.itcast.cn" target="\_blank">  <span class="s\_name">北京</span>  </a>  </li>  <li>  <a href="http://sh.itcast.cn" target="\_blank">  <span class="s\_name">上海</span>  </a>  </li>  <li>  <a href="http://gz.itcast.cn" target="\_blank">  <span abc="123" class="s\_name">广州</span>  </a>  </li>  <ul>  <li>天津</li>  </ul>  </ul>  </div>  </div>  </div>  </body>  </html> |

@Test  
public void tesString() throws Exception{  
 *//解析字符串* String content = FileUtils.*readFileToString*(new File("F:\\project\\IDEA\\study\\javaee57\\web-crawler-first\\docs\\test.html"), "utf8");  
 System.*out*.println(content);  
  
 *//第一参数：访问的url,第二个参数是访问时的超时时间* Document doc = Jsoup.*parse*(content);  
  
  
 *//使用标签选择器，获取title中的内容* String title = doc.getElementsByTag("title").first().text();  
 System.*out*.println(title);  
}

#### 4.2.3 解析文件

Jsoup可以直接解析文件，并封装为Document对象

@Test  
 public void testFile() throws Exception{  
 *//第一参数：访问的url,第二个参数是访问时的超时时间* Document doc = Jsoup.*parse*(new File("F:\\project\\IDEA\\study\\javaee57\\web-crawler-first\\docs\\test.html"),"utf8");  
  
  
 *//使用标签选择器，获取title中的内容* String title = doc.getElementsByTag("title").first().text();  
 System.*out*.println(title);  
 }  
}

#### 4.2.4 使用dom方式

元素获取

1. 根据id查询元素getElementById
2. 根据标签获取元素getElementsByTag
3. 根据class获取元素getElementsByClass
4. 根据属性获取元素getElementsByAttribute

@Test  
public void testDOM() throws Exception{  
 *//第一参数：访问的url,第二个参数是访问时的超时时间* Document doc = Jsoup.*parse*(new File("F:\\project\\IDEA\\study\\javaee57\\web-crawler-first\\docs\\test.html"),"utf8");  
  
 *//1.根据id查询元素getElementById* String city\_bj = doc.getElementById("city\_bj").text();  
 System.*out*.println(city\_bj);  
 *//2.根据标签获取元素getElementsByTag* String span = doc.getElementsByTag("span").first().text();  
 System.*out*.println(span);  
 *//3.根据class获取元素getElementsByClass* String class\_a\_class\_b = doc.getElementsByClass("class\_a class\_b").first().text();  
 String class\_a = doc.getElementsByClass("class\_a").first().text();  
 System.*out*.println(class\_a\_class\_b);  
 System.*out*.println(class\_a);  
 *//4.根据属性获取元素getElementsByAttribute* String abc = doc.getElementsByAttribute("abc").first().text();  
 System.*out*.println(abc);  
  
 *//5.组合* String abc1 = doc.getElementsByAttributeValue("href","http://gz.itcast.cn").first().text();  
 System.*out*.println(abc1);  
  
}

元素中获取数据

1. 从元素中获取id
2. 从元素中获取className
3. 从元素中获取属性的值attr
4. 从元素中获取所有属性attributes
5. 从元素中获取文本内容text

@Test  
 public void testData() throws Exception{  
 *//第一参数：访问的url,第二个参数是访问时的超时时间* Document doc = Jsoup.*parse*(new File("F:\\project\\IDEA\\study\\javaee57\\web-crawler-first\\docs\\test.html"),"utf8");  
  
 Element element = doc.getElementById("test");  
  
 String str = "";  
  
*// str = element.id();  
// str = element.className();  
// Set<String> classNames = element.classNames();  
// for (String className : classNames) {  
// System.out.println(className);  
// }  
  
// str = element.attr("id");* Attributes attributes = element.attributes();  
  
 System.*out*.println(str);  
  
 *//使用标签选择器，获取title中的内容* String title = doc.getElementsByTag("title").first().text();  
 System.*out*.println(title);  
 }

#### 4.2.5 使用选择器

jsoup elements对象支持类似于CSS (或jquery)的选择器语法，来实现非常强大和灵活的查找功能。这个select 方法在Document, Element,或Elements对象中都可以使用。且是上下文相关的，因此可实现指定元素的过滤，或者链式选择访问。

Select方法将返回一个Elements集合，并提供一组方法来抽取和处理结果。

**tagname**: 通过标签查找元素，比如：span

**#id**: 通过ID查找元素，比如：# city\_bj

**.class**: 通过class名称查找元素，比如：.class\_a

**[attribute]**: 利用属性查找元素，比如：[abc]

**[attr=value]**: 利用属性值来查找元素，比如：[class=s\_name]

#### 4.2.6 selector选择器概述

@Test  
public void tesSelector() throws Exception{  
 *//第一参数：访问的url,第二个参数是访问时的超时时间* Document doc = Jsoup.*parse*(new File("F:\\project\\IDEA\\study\\javaee57\\web-crawler-first\\docs\\test.html"),"utf8");  
  
 *//标签获取元素* Elements elements = doc.select("span");  
  
 for (Element element : elements) {  
 System.*out*.println(element.text());  
 }  
  
 *//id获取元素* String city\_bj = doc.select("#city\_bj").first().text();  
 System.*out*.println(city\_bj);  
  
 *//class获取元素* String class\_a\_class\_b = doc.select(".class\_a").first().text();  
 System.*out*.println(class\_a\_class\_b);  
  
  
 *//属性获取元素* String abd = doc.select("[abc]").first().text();  
 System.*out*.println(abd);  
}

#### 4.2.7 selector选择器组合使用

**el#id**: 元素+ID，比如： h3#city\_bj

**el.class**: 元素+class，比如： li.class\_a

**el[attr]**: 元素+属性名，比如： span[abc]

**任意组合**: 比如：span[abc].s\_name

**ancestor child**: 查找某个元素下子元素，比如：.city\_con li 查找"city\_con"下的所有li

**parent > child**: 查找某个父元素下的直接子元素，比如：

.city\_con > ul > li 查找city\_con第一级（直接子元素）的ul，再找所有ul下的第一级li

**parent > \***: 查找某个父元素下所有直接子元素

@Test  
public void tesSelectorAll() throws Exception{  
 *//第一参数：访问的url,第二个参数是访问时的超时时间* Document doc = Jsoup.*parse*(new File("F:\\project\\IDEA\\study\\javaee57\\web-crawler-first\\docs\\test.html"),"utf8");  
  
 *//el#id: 元素+ID，比如： h3#city\_bj* Elements elements = doc.select("h3#city\_bj");  
  
 for (Element element : elements) {  
 System.*out*.println(element.text());  
 }  
  
 *//el.class: 元素+class，比如： li.class\_a* String class\_a\_class\_b = doc.select("li.class\_a").first().text();  
 System.*out*.println(class\_a\_class\_b);  
  
  
 *//el[attr]: 元素+属性名，比如： span[abc]* String abd = doc.select("span[abc]").first().text();  
 System.*out*.println(abd);  
  
 *//任意组合: 比如：span[abc].s\_name* String na = doc.select("span[abc].s\_name").first().text();  
 System.*out*.println(na);  
  
 *//ancestor child: 查找某个元素下子元素，比如：.city\_con li 查找"city\_con"下的所有li* Elements elements1 = doc.select(".city\_con li");  
 for (Element element : elements1) {  
 System.*out*.println(element.text());  
 }  
  
 *//parent > child: 查找某个父元素下的直接子元素，比如：  
 //.city\_con > ul > li 查找city\_con第一级（直接子元素）的ul，再找所有ul下的第一级li* String na1 = doc.select(".city\_con > ul > li").first().text();  
 System.*out*.println(na1);  
  
 *//parent > \*: 查找某个父元素下所有直接子元素* String na3 = doc.select(".city\_con > \*").first().text();  
 System.*out*.println(na3);  
  
}

## 第五章 爬虫案例

学习了HttpClient和Jsoup，就掌握了如何抓取数据和如何解析数据，接下来，我们做一个小练习，

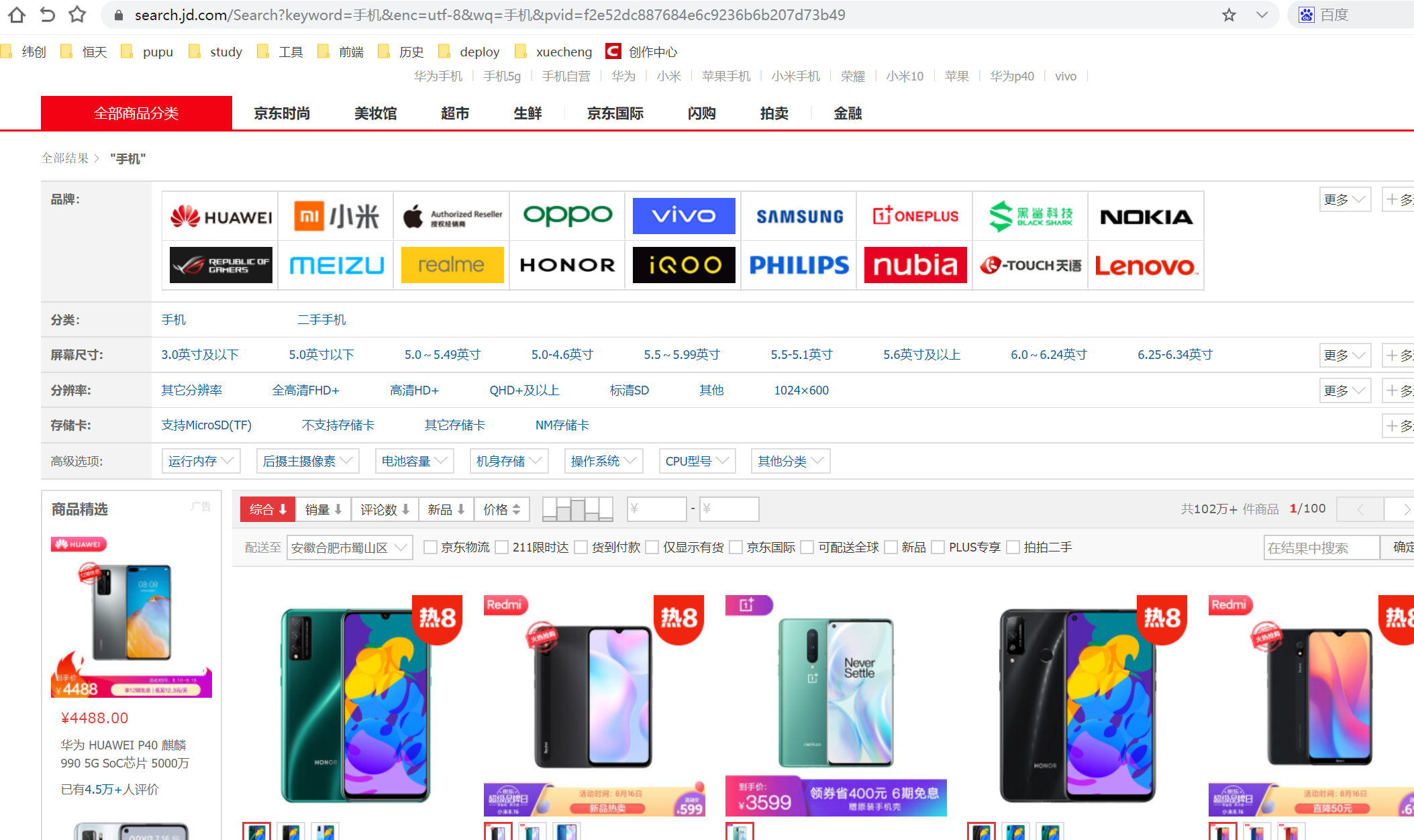
把京东的手机数据抓取下来。

主要目的是HttpClient和Jsoup的学习。

### 5.1 需求分析

首先访问京东，搜索手机，分析页面，我们抓取以下商品数据：

商品图片、价格、标题、商品详情页



#### 5.1.1 **SPU和SKU**

除了以上四个属性以外，我们发现上图中的苹果手机有四种产品，我们应该每一种都要抓取。那么这里就必须要了解spu和sku的概念。

**SPU** = Standard Product Unit （标准产品单位）

SPU是商品信息聚合的最小单位，是一组可复用、易检索的标准化信息的集合，该集合描述了一个产品的特性。通俗点讲，属性值、特性相同的商品就可以称为一个SPU。

例如上图中的苹果手机就是SPU，包括红色、深灰色、金色、银色

**SKU**=stock keeping unit(库存量单位)

SKU即库存进出计量的单位， 可以是以件、盒、托盘等为单位。SKU是物理上不可分割的最小存货单元。在使用时要根据不同业态，不同管理模式来处理。在服装、鞋类商品中使用最多最普遍。

例如上图中的苹果手机有几个款式，红色苹果手机，就是一个sku

查看页面的源码也可以看出区别



### 5.2 开发准备

#### 6.2.1 **数据库表分析**

根据需求分析，我们创建的表如下：

CREATE TABLE `jd\_item` (

`id` bigint(10) NOT NULL AUTO\_INCREMENT COMMENT '主键id',

`spu` bigint(15) DEFAULT NULL COMMENT '商品集合id',

`sku` bigint(15) DEFAULT NULL COMMENT '商品最小品类单元id',

`title` varchar(100) DEFAULT NULL COMMENT '商品标题',

`price` bigint(10) DEFAULT NULL COMMENT '商品价格',

`pic` varchar(200) DEFAULT NULL COMMENT '商品图片',

`url` varchar(200) DEFAULT NULL COMMENT '商品详情地址',

`created` datetime DEFAULT NULL COMMENT '创建时间',

`updated` datetime DEFAULT NULL COMMENT '更新时间',

PRIMARY KEY (`id`),

KEY `sku` (`sku`) USING BTREE

) ENGINE=InnoDB AUTO\_INCREMENT=1 DEFAULT CHARSET=utf8 COMMENT='京东商品表';

#### 6.2.2 **添加依赖**

使用Spring Boot+Spring Data JPA和定时任务进行开发，

需要创建Maven工程并添加以下依赖

<?xml version="1.0" encoding="UTF-8"?>  
<project xmlns="http://maven.apache.org/POM/4.0.0"  
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"  
 xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/xsd/maven-4.0.0.xsd">  
 <modelVersion>4.0.0</modelVersion>  
 <parent>  
 <groupId>org.springframework.boot</groupId>  
 <artifactId>spring-boot-starter-parent</artifactId>  
 <version>2.0.2.RELEASE</version>  
 </parent>  
 <groupId>cn.itcast.crawler</groupId>  
 <artifactId>itcast-crawler-jd</artifactId>  
 <version>1.0-SNAPSHOT</version>  
  
 <dependencies>  
 <!--SpringMVC-->  
 <dependency>  
 <groupId>org.springframework.boot</groupId>  
 <artifactId>spring-boot-starter-web</artifactId>  
 </dependency>  
  
 <!--SpringData Jpa-->  
 <dependency>  
 <groupId>org.springframework.boot</groupId>  
 <artifactId>spring-boot-starter-data-jpa</artifactId>  
 </dependency>  
  
 <!--MySQL连接包-->  
 <dependency>  
 <groupId>mysql</groupId>  
 <artifactId>mysql-connector-java</artifactId>  
 </dependency>  
  
 <!-- HttpClient -->  
 <dependency>  
 <groupId>org.apache.httpcomponents</groupId>  
 <artifactId>httpclient</artifactId>  
 </dependency>  
  
 <!--Jsoup-->  
 <dependency>  
 <groupId>org.jsoup</groupId>  
 <artifactId>jsoup</artifactId>  
 <version>1.10.3</version>  
 </dependency>  
  
 <!--工具包-->  
 <dependency>  
 <groupId>org.apache.commons</groupId>  
 <artifactId>commons-lang3</artifactId>  
 </dependency>  
 </dependencies>  
</project>

#### 6.2.3 添加配置文件

加入application.properties配置文件

#DB Configuration:  
spring.datasource.driverClassName=com.mysql.jdbc.Driver  
spring.datasource.url=jdbc:mysql://127.0.0.1:3306/crawler  
spring.datasource.username=root  
spring.datasource.password=root  
  
#JPA Configuration:  
spring.jpa.database=*MySQL*spring.jpa.show-sql=**true**

### 5.3 代码实现

#### 5.3.1 **编写pojo**

#### 5.3.2 **编写dao**

#### 5.3.3 **编写service**

#### 5.3.4 **编写引导类**

#### 5.3.5 **封装HttpClient**

#### 5.3.6 **实现数据抓取**

# 网络爬虫第二天

## 第六章 Kafka监控

## 第六章 Flume对接Kafka

## 第七章 Kafka面试题

1、Kafka中的ISP（InSyncRepli）、OSR(OutSyncRepli)、AR(AllRepli)代表什么？

2、Kafka中的HW、LEO等分别代表什么？

3、Kafka中怎么体现消息顺序性的？

4、Kafka中的分区器、序列化器、拦截器是否了解？它们之间的处理顺序是什么？

5、Kafka生产者客户端的整体结构什么样子的？使用了几个线程来处理？分别是什么？

6、“消费者组中的消费者个数如果超过topic的分区，那么就会有消费者消费不到数据”这句话是否正确？

7、消费者提交消费位移时提交的当前消费到的最新消息的offset还是offset+1？

8、有哪些情形会造成重复消费？先消费再set

9、哪些情景会造成消息漏消费？先set再消费

10、当你使用kafka-topic.sh创建（删除）了一个topic之后，Kafka背后会执行什么逻辑？

1）在zookeeper中的/brokers/topics节点下创建一个新的topic节点，如：/brokers/topics/first

2）触发Controller的监听程序

3）kafka Controller负责topic的创建工作，并更新metadata cache

11、topic的分区数可不可以增加？如果可以怎么增加？如果不可以又是为什么？

12、topic的分区数可不可以减少？如果可以怎么减少？如果不可以又是为什么？

13、kafka有内部的topic吗？如果有是什么？有什么作用？存offset

14、kafka分区分配的概念？

15、简述kafka的日志目录结构；index+log

16、如果制指定了一个offset，kafka Controller怎么查找到对应的消息？

17、聊一聊kafka Controller的作用？

18、kafka中有哪些地方需要选举？这些地方的选举策略又有哪些？controller leader

19、失效副本是指什么？有哪些应对措施？

20、kafka的哪些设计让他有如此高的性能？分布式、顺序写磁盘、0拷贝