**网络爬虫**

**第一天**

# 课程计划

1. 入门程序
2. 网络爬虫介绍
3. HttpClient抓取数据
4. Jsoup解析数据
5. 爬虫案例

# 网络爬虫

网络爬虫（Web crawler），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

## 爬虫入门程序

### 环境准备

* JDK1.8
* IntelliJ IDEA
* IDEA自带的Maven

### 环境准备

创建Maven工程itcast-crawler-first并给pom.xml加入依赖

<dependencies>  
 <!-- HttpClient -->  
 <dependency>  
 <groupId>org.apache.httpcomponents</groupId>  
 <artifactId>httpclient</artifactId>  
 <version>4.5.3</version>  
 </dependency>  
  
 <!-- 日志 -->  
 <dependency>  
 <groupId>org.slf4j</groupId>  
 <artifactId>slf4j-log4j12</artifactId>  
 <version>1.7.25</version>  
 </dependency>  
</dependencies>

### 加入log4j.properties

log4j.rootLogger=DEBUG,A1  
log4j.logger.cn.itcast = DEBUG  
  
log4j.appender.A1=org.apache.log4j.ConsoleAppender  
log4j.appender.A1.layout=org.apache.log4j.PatternLayout  
log4j.appender.A1.layout.ConversionPattern=%-d{yyyy-MM-dd HH:mm:ss,SSS} [%t] [%c]-[%p] %m%n

### 编写代码

编写最简单的爬虫，抓取传智播客首页：http://www.itcast.cn/

public static void main(String[] args) throws Exception {  
 CloseableHttpClient httpClient = HttpClients.*createDefault*();  
  
 HttpGet httpGet = new HttpGet("http://www.itcast.cn/");  
  
 CloseableHttpResponse response = httpClient.execute(httpGet);  
  
 if (response.getStatusLine().getStatusCode() == 200) {  
 String content = EntityUtils.*toString*(response.getEntity(), "UTF-8");  
 System.*out*.println(content);  
 }  
}

测试结果：可以获取到页面数据

# 网络爬虫

## 网络爬虫介绍

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。

## 为什么学网络爬虫

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。

* + - 1. 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。

* + - 1. 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。

* + - 1. 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。