**2. 快速开始**

WebMagic主要包含两个jar包：webmagic-core-{version}.jar和webmagic-extension-{version}.jar。在项目中添加这两个包的依赖，即可使用WebMagic。

WebMagic默认使用Maven管理依赖，但是你也可以不依赖Maven进行使用。

### 2.1 使用Maven

WebMagic基于Maven进行构建，推荐使用Maven来安装WebMagic。在你自己的项目（已有项目或者新建一个）中添加以下坐标即可：

<dependency>

<groupId>us.codecraft</groupId>

<artifactId>webmagic-core</artifactId>

<version>0.7.3</version>

</dependency>

<dependency>

<groupId>us.codecraft</groupId>

<artifactId>webmagic-extension</artifactId>

<version>0.7.3</version>

</dependency>

如果你对Maven使用还不熟悉，推荐看看@黄勇的博客：[《Maven 那点事儿》](http://my.oschina.net/huangyong/blog/194583)。

WebMagic使用slf4j-log4j12作为slf4j的实现.如果你自己定制了slf4j的实现，请在项目中去掉此依赖。

<dependency>

<groupId>us.codecraft</groupId>

<artifactId>webmagic-extension</artifactId>

<version>0.7.3</version>

<exclusions>

<exclusion>

<groupId>org.slf4j</groupId>

<artifactId>slf4j-log4j12</artifactId>

</exclusion>

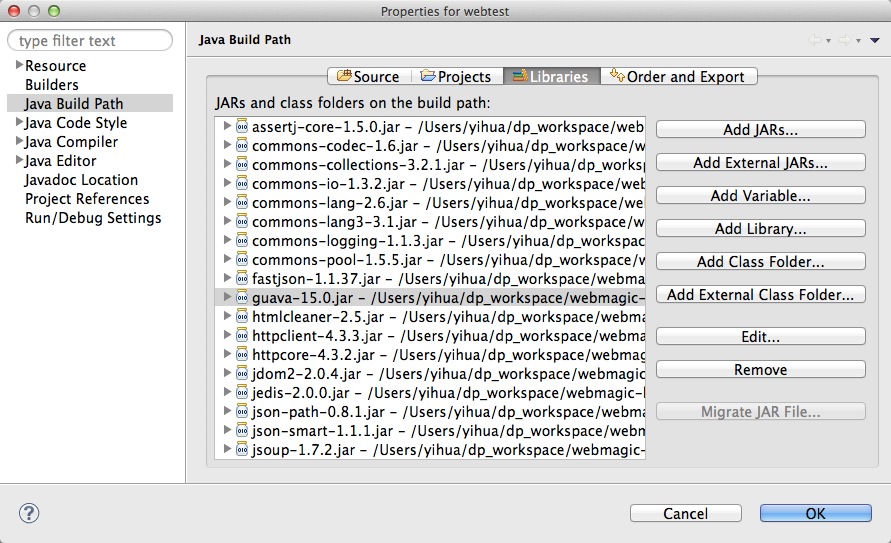
</exclusions>

</dependency>

### 2.2 不使用Maven

不使用maven的用户，可以去[http://webmagic.io](http://webmagic.io/)中下载最新的jar包。我将所有依赖的jar都打包在了一起，点[这里](http://webmagic.io/download.html)下载。

下载之后进行解压，然后在项目中import即可。



因为WebMagic提倡自己定制，所以项目的源码还是有必要看的。在[http://webmagic.io](http://webmagic.io/)上，你可以下载最新的webmagic-core-{version}-sources.jar和webmagic-extension-{version}-sources.jar，点击"Attach Source"即可。

### 2.3 第一个爬虫项目

在你的项目中添加了WebMagic的依赖之后，即可开始第一个爬虫的开发了！我们这里拿一个抓取Github信息的例子：

import us.codecraft.webmagic.Page;

import us.codecraft.webmagic.Site;

import us.codecraft.webmagic.Spider;

import us.codecraft.webmagic.processor.PageProcessor;

public class GithubRepoPageProcessor implements PageProcessor {

//me:站点抓取配置

private Site site = Site.me().setRetryTimes(3).setSleepTime(100)

//me：实现接口方法

@Override

public void process(Page page) {

//me：添加新的链接 page.addTargetRequests(page.getHtml().links().regex("([https://github\\.com/\\w+/\\w+)").all())](https://github\\.com/\\w+/\\w+)%22).all()));

//me：提取信息

page.putField("author", page.getUrl().regex("https://github\\.com/(\\w+)/.\*").toString());

page.putField("name", page.getHtml().xpath("//h1[@class='entry-title public']/strong/a/text()").toString());

if (page.getResultItems().get("name")==null){

//skip this page

page.setSkip(true);

}

page.putField("readme", page.getHtml().xpath("//div[@id='readme']/tidyText()"));

}

@Override

public Site getSite() {

return site;

}

public static void main(String[] args) {

Spider.create(new GithubRepoPageProcessor()).addUrl("https://github.com/code4craft").thread(5).run();

}

}

点击main方法，选择“运行”，你会发现爬虫已经可以正常工作了！

