Atitit.数据采集器 dataspider

/atiplat\_cms/src/com/attilax/WebInfoX.java @dep

<http://cl.cmcher.com/thread0806.php?fid=16&search=&page=2>

/atiplat\_cms/src/com/attilax/dataspider/TsaolyoNetDataSpider.java

# 爬虫注意事项

设置useragent as ff

https的注意。。

Timeout http打阻塞。。一定要timeout

Pics 多线程，一定要。

日志要指明 p1art1

主要是因为**[Java](http://lib.csdn.net/base/17" \o "Java EE知识库" \t "http://blog.csdn.net/witsmakemen/article/details/_blank)**自己的HttpURLConnection对SSL支持的不好，而且控制起来不太方便，而且HttpClient还支持抓取非信任的站点，别的实现方式貌似需要在代码中显式导入证书。

作者:: 绰号:老哇的爪子 （ 全名：：Attilax Akbar Al Rapanui 阿提拉克斯 阿克巴 阿尔 拉帕努伊 ） 汉字名：艾提拉（艾龙），  EMAIL:1466519819@qq.com

转载请注明来源： http://blog.csdn.net/attilax

# 需要override打三个函数

**public** List<String> **getpageUrls**()

**public** List **getArtListByPagehtml**(**String** html) {

**public** List<String> **getPics\_byHtml**

# 使用方法与参数

**TsaolyoNetDataSpider** **x** = **new** TsaolyoNetDataSpider();

// x.fileName=args[0];// "c:\\r2.csv";

x.picSaveDir = "c:\\0picSaveDir";

x.startPage = **Integer**.*parseInt*(**System**.*getProperty*("startPage", "1"));

x.endPage = **Integer**.*parseInt*(**System**.*getProperty*("endPage"));

;

x.exec();