Heritrix 网络爬虫文档

1. 项目源码仓库: <https://github.com/mydipcom/crawler.git>
2. 目前已经实现的功能

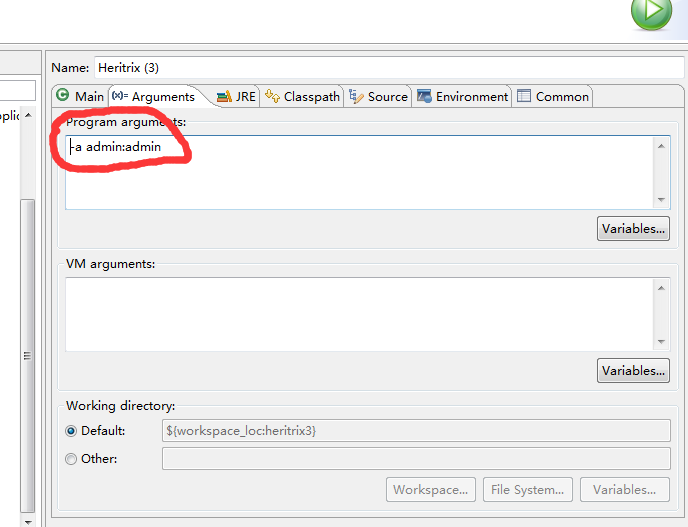
A、netporter、mrorter、米兰站、苏宁产品信息的抓取和写入本地数据库.

B、实现对netporter下架商品自动下架（接口方式实现）

三、项目启动

A、将项目导入eclipse

B、启动：Java application 方式运行org.archive.crawler.Heritrix类



-a admin:admin 设置浏览器端登录用户名密码

1. 运行爬虫任务

浏览器访问地址：<https://localhost:8443>



以nap为例运行任务



Build : 建立任务

Launch：启动任务

Pause：暂停

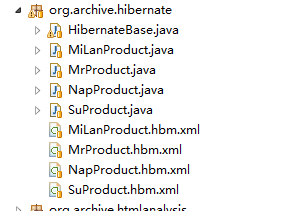
Unpause：恢复

Checkpoint:添加检查点

运行步骤：

首次启动：build 🡪launch🡪Checkpoint🡪unpause

1. 自定义代码

实体：

解析代码：

