# 通用爬虫nutch

#### 目的

使用最新的nutch，抓取https://www.frbao.com/ 网站的网页，存入hbase中，使用solr索引网页

#### 软件准备

* 下载nutch的最新源码http://svn.apache.org/repos/asf/nutch/branches/2.x
* 下载hbase的linux 0.94最后的版本；
* 下载solr4的windows最后版本；

#### 软件安装配置

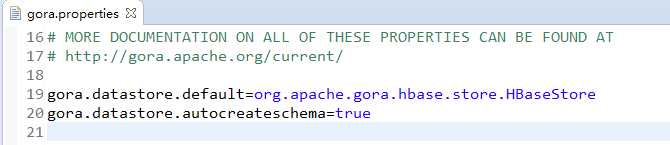
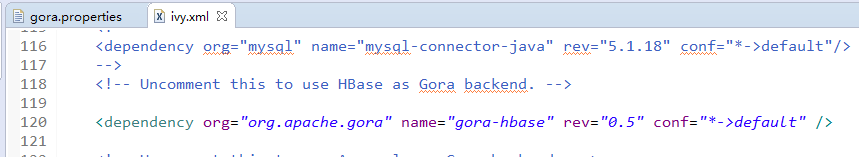
1. **hbase：**

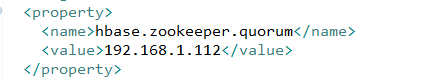
在linux中以Standalone 方式安装在local filesystem

1. **solr4**

直接解压， example目录下运行java -jar start.jar

1. **nutch2**

源码导入eclipse中，在ivy.xml和gora.properties中开启gora对hbase的支持，  


在hbase-site.xml中指定zookeeper的地址，  


在nutch-site.xml中添加

执行build.xml中的ant任务，编译项目，下载依赖包；

<property>

<name>http.agent.name</name>

<value>Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/45.0.2454.7 Safari/537.36 nutch/2.4</value>

</property>

<property>

<name>storage.data.store.class</name>

<value>org.apache.gora.hbase.store.HBaseStore</value>

</property>

#### 在eclipse中执行nutch，抓取网页

1. **准备种子文本文件**

把https://www.frbao.com/写入文件中

1. **eclipse中执行nutch的多个任务**

inject org.apache.nutch.crawl.InjectorJob

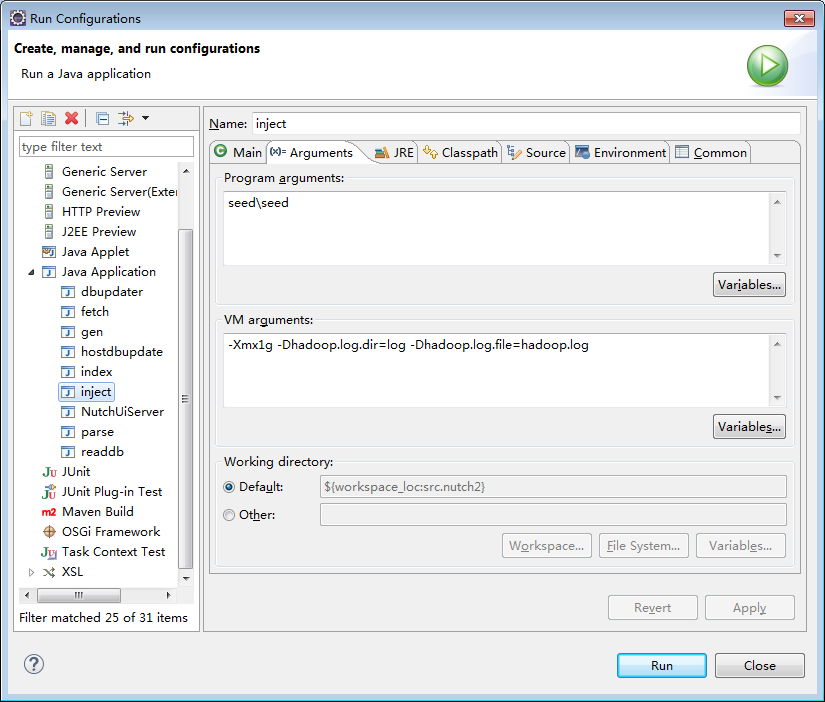
generate org.apache.nutch.crawl.GeneratorJob

fetch org.apache.nutch.fetcher.FetcherJob

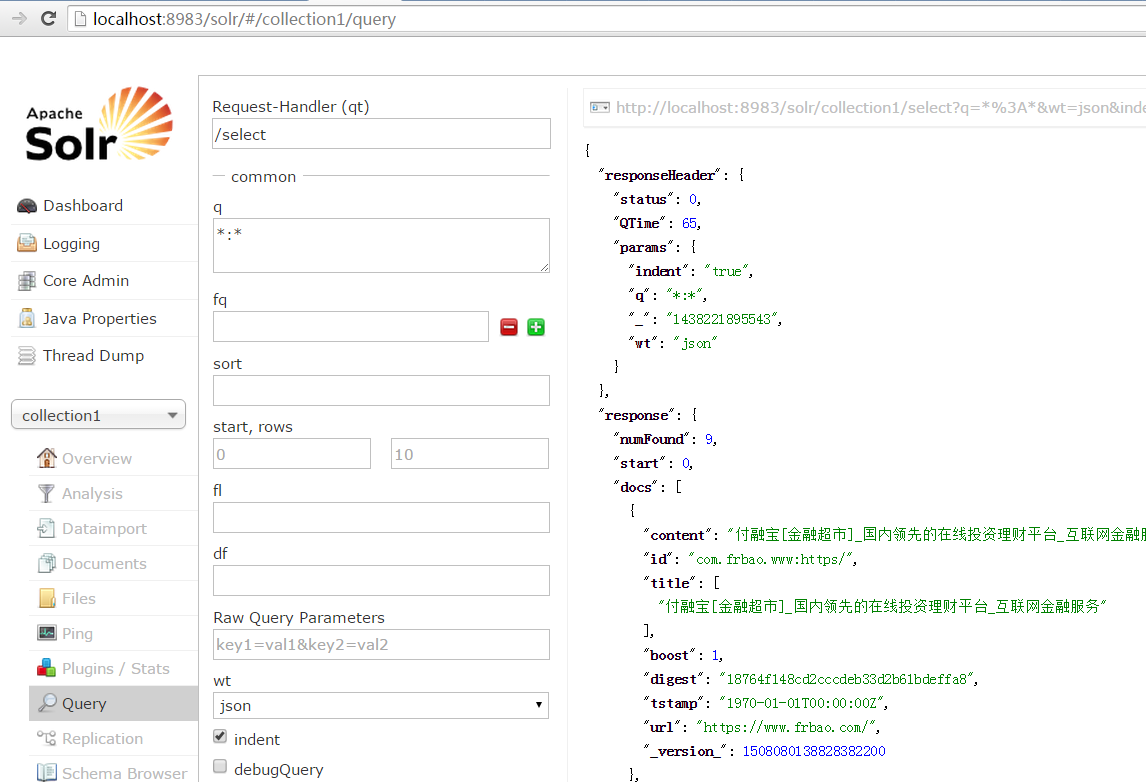
parse org.apache.nutch.parse.ParserJob

updatedb org.apache.nutch.crawl.DbUpdaterJob

index org.apache.nutch.indexer.IndexingJob



#### solr中查看索引的结果



# 自建程序抓取解析baidu的新闻

#### 目的

抓取百度新闻的索引页，抽取新闻的链接；

抓取新闻的内容页，提取正文；

#### 软件准备

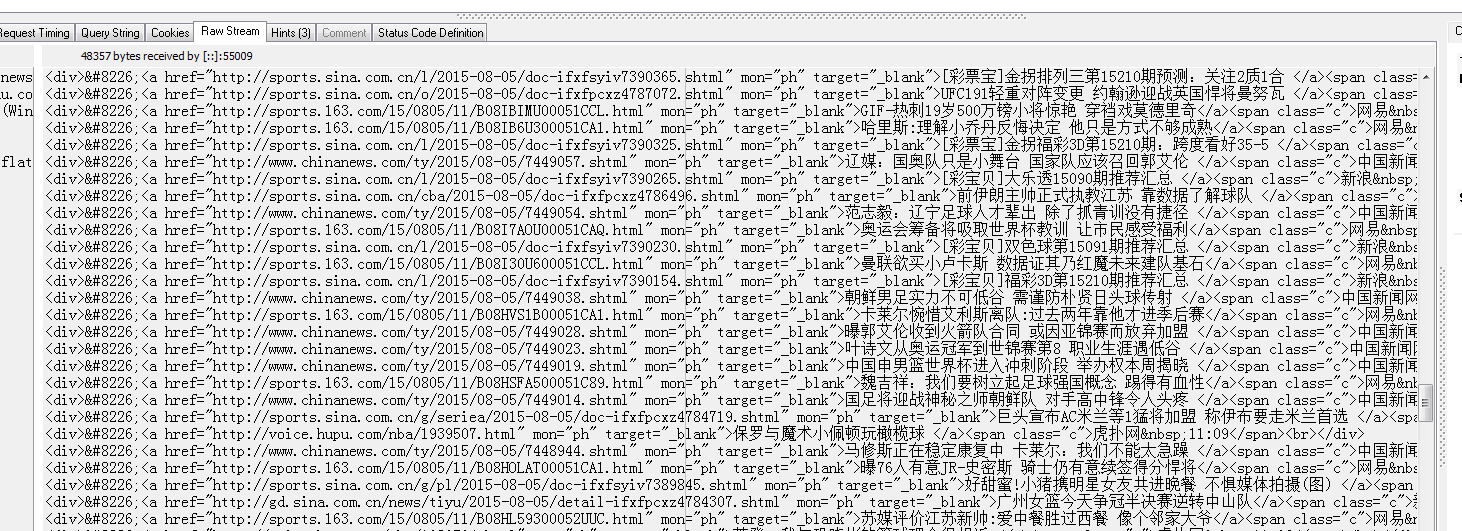
* httpclient4.3及以上版本；

#### 程序实现

1. **抓取网页：**



1. **正则表达式抽取索引页的链接：**



"&#8226;<a href=\"(.\*?)\""

1. **提取网页正文：**

* 基于标签用途的正文提取算法（比如title或h1,h2标签一般用作标题，p一般表示正文段落，根据标签的含义去提取正文）
* 基于标签密度判定（这个简单，说白了就是字符统计，正文部分html标签的密度比较低，确定一个阈值，按照标签密度提取正文部分）
* 基于数据挖掘思想的网页正文抽取方法（该方法将网页源代码进行线性化重构,然后利用重构后的代码进行网页噪声的初步去除,再经过文本分类、聚类得到网页正文的脉络段落,最后通过吸收伪噪声段落生成网页正文。）

