基本配置

* jdbc数据库
* 文件下载，文字识别的线程数以及线程爬取的间隔
* 任务周期
* 文件保存目录

代理的配置

* <host>主机名
* <port>端口号
* <auth>用户名密码(如果需要的话)

数据源配置

* 线程数、线程爬取的间隔、是否使用代理
* 数据源种子

<seed url=*"http://www.okbuy.com/product/filter?category=54&#38;per\_page=\*00"* from=*"1"* to=*"1"*

query=*"li.pc\_list\_hover>div.img>a"* attr=*"href"*>

<columns>

<column name=*"style"* value=*"运动鞋"* type=*"VARCHAR(100)"* />

</columns>

</seed>

1. 种子url，\*代表页数，from-to是从第几页爬到第几页
2. Query爬取数据url的规则（jquery规则）
3. Attr拿该标签的attr属性内容，如果为空，则拿该标签的文本内容
4. Columns这个种子能确定的某些字段

* 数据源爬取规则

<columns>

<column name=*"id"* type=*"INT(11)"*>

<rule type=*"html"* value=*"ul.pProductTitle>li>p"* index=*"0"* />

</column>

<column name=*"brand"* type=*"INT(11)"*>

<rule type=*"text"* url=*" http://www.*okbuy*.com/product/ajax\_product\_info/{id}"* value=*"0"* />

</column>

<column name=*"arno"* type=*"VARCHAR(100)"*>

<rule type=*"json"* url=*"http://www.*okbuy*.com/product/ajax\_product\_info/{id}"* value=*"/o/salepr"* index=*"0"* />

</column>

<column name=*"price"* type=*"DOUBLE"* ocr\_column=*"price\_ocr\_url"*>

<rule type=*"html"* value=*"ul#summary>li#summary-price>div.dd>strong>img"* index=*"0"* attr=*"src"*/>

</column>

<column name=*"img\_1"* type=*"VARCHAR(1000)"* file\_column=*"img\_1\_url"*

file\_path=*"{yyyy}/{MM}/"* >

<rule type=*"html"* value=*"table.multi\_view>tbody>tr>td>a"* index=*"0"* attr=*"href"*/>

</column>

</columns>

1. Column:字段

如果seed.column里有的字段，最好不要再出现。（seed里有的字段，说明这个字段值已经在确定了，不需要再抽取）

1. Name:字段名

唯一的，一张表可能有两个字段名字相同吗？

1. Type:字段类型
2. Ocr\_column:需要识别的图片url字段

需要识别的图片的url字段

1. File\_column:需要下载的文件url字段

需要下载的文件的url字段

1. File\_path:下载的相对路径

完整路径：basepath+数据源+file\_path

1. Overwrite:要是该字段已经有值了，是否覆盖。默认是false
2. Rule.type:爬取的类型
3. Html query规则抽取
4. Json json规则抽取
5. Xml xml规则抽取
6. Text 正则规则抽取
7. Rule:value:爬取的规则
8. Html对应query规则
9. Json、xml对应xpath规则
10. Text对应正则
11. Rule:url:异步请求的url

如果是json、xml规则，rule.url不能为空。{name}对应url里的参数，name必须有值，name就是字段名

1. Rule:attr: 拿该标签的attr属性内容，如果为空，则拿该标签的文本内容
2. Rule:index:拿第几个标签,默认是0（第一个）