**configs详解——之field**

field定义一个抽取项, 一个field可以定义下面这些东西

**name**

给此项数据起个变量名  
变量名中不能包含.  
如果抓取到的数据想要以文章或者问答的形式发布到网站(WeCenter,  
WordPress, Discuz!等), field的命名请参考两个完整demo中的命名, 否则无法发布成功

**String类型 不能为空**

举个栗子:

给field起了个名字叫content

array(

'name' => "content",

'selector' => "//\*[@id='single-next-link']"

)

**selector**

定义抽取规则, 默认使用xpath  
如果使用其他类型的, 需要指定selector\_type

**String类型 不能为空**

举个栗子:  
使用xpath来抽取糗事百科的笑话内容，selector的值就是内容的xpath

array(

'name' => "content",

'selector' => "//\*[@id='single-next-link']"

)

**selector\_type**

抽取规则的类型

目前可用[xpath](http://www.w3school.com.cn/xpath/index.asp), [jsonpath](http://www.cnblogs.com/draem0507/p/5111002.html), [regex](http://www.runoob.com/regexp/regexp-tutorial.html)  
默认xpath

**枚举类型**

栗子1:  
selector默认使用xpath

array(

'name' => "content",

'selector' => "//\*[@id='single-next-link']" // xpath抽取规则

)

栗子2:  
使用正则表达式来抽取数据

array(

'name' => "content",

'selector\_type' => 'regex',

'selector' => '#<div\sclass="content">([^/]+)</div>#i' // regex抽取规则

)

**required**

定义该field的值是否必须, 默认false  
赋值为true的话, 如果该field没有抽取到内容, 该field对应的整条数据都将被丢弃

**布尔类型**

举个栗子:

array(

'name' => "content",

'selector' => "//\*[@id='single-next-link']",

'required' => true

)

**repeated**

定义该field抽取到的内容是否是有多项, 默认false  
赋值为true的话, 无论该field是否真的是有多项, 抽取到的结果都是数组结构

**布尔类型**

举个栗子:  
爬取的网页中包含多条评论，所以抽取评论的时候要将repeated赋值为true

array(

'name' => "comments",

'selector' => "//\*[@id='zh-single-question-page']//a[contains(@class,'zm-item-tag')]",

'repeated' => true

)

**children**

为此field定义子项  
子项的定义仍然是一个fields数组  
没错, 这是一个树形结构

**数组类型**

举个栗子:  
抓取糗事百科的评论，每个评论爬取了内容，点赞数

array(

'name' => "article\_comments",

'selector' => "//div[contains(@class,'comments-wrap')]",

'children' => array(

array(

'name' => "replay",

'selector' => "//div[contains(@class,'replay')]",

'repeated' => true,

),

array(

'name' => "report",

'selector' => "//div[contains(@class,'report')]",

'repeated' => true,

)

)

)

**source\_type**

该field的数据源, 默认从当前的网页中抽取数据  
选择attached\_url可以发起一个新的请求, 然后从请求返回的数据中抽取  
选择url\_context可以从当前网页的url附加数据（点此查看“url附加数据”实例解析）中抽取

**枚举类型**

**attached\_url**

当source\_type设置为attached\_url时, 定义新请求的url

**String类型**

举个栗子:  
当爬取的网页中某些内容需要异步加载请求时，就需要使用attached\_url，比如，抓取知乎回答中的评论部分，就是通过AJAX异步请求的数据

array(

'name' => "comment\_id",

'selector' => "//div/@data-aid",

),

array(

'name' => "comments",

'source\_type' => 'attached\_url',

// "comments"是从发送"attached\_url"这个异步请求返回的数据中抽取的

// "attachedUrl"支持引用上下文中的抓取到的"field", 这里就引用了上面抓取的"comment\_id"

'attached\_url' => "https://www.zhihu.com/r/answers/{comment\_id}/comments",

'selector\_type' => 'jsonpath'

'selector' => "$.data",

'repeated => true,

'children' => array(

...

)

}