爬虫基本规则整理

1．静态网页

```
url: http://cqfb.people.com.cn/

regulation: {
        "url_xpath": "//div[@class='news1']/h4/a",
        "title_xpath": ""
        }
```

首先判断字典里面key为url的值是否为空，空和不为空分开处理。为的是title不包含链接的特殊网站。

2．api

普通的没有特殊字符串的：
比如返回格式为：

```
 {
   'msg': '',
   'status': '',
   'date': '',
   'page_num': '',
   'data':[
       {
         'title': '我爱中国',
         'url': 'www.aizhongguo.com'
       },
       '''
   ]
 }
```

这种情况我们要取的数据在一个字典里面的data里面，然后data里面是很多列表，列表里面是我们要的每一条数据，解析规则为：

```
  {
   'list_path':[0,data]
   'url_path':['data_field1',...,url],
   'title_path':['data_field1,...,title],
   'preprocess': {'type':'None'}}
   }
```

如果我们要的data列表在字典里面的多层嵌套，只需要一次将嵌套的key添加到一次加到regulation的列表里面。

首尾有特殊字符串的：

```
比如返回格式为：
 HSHFJKSHD data = (
    {
   'msg': '',
   'status': '',
   'date': '',
   'page_num': '',
   'data':[
       {
         'title': '我爱中国',
         'url': 'www.aizhongguo.com'
       },
       '''
   ]
 }
 )

规则为:
  {
   'list_path':[0,data]
   'url_path':['data_field1',...,url],
   'title_path':['data_field1,...,title],
   'preprocess': {'type':'TRIM', 'meta'':{'index': (18,-1)}}
   }

示例网站：
http://qc.wa.news.cn/nodeart/list?nid=11100290&pgnum=1&cnt=50&attr=&tp=1&orderby=1&callback=jQuery171000300873536803036 86_1568706392177&_=1568706400699
对应示例规则：
{
"list_path":["data"],
"url_path": ["LinkUrl"],
 "title_path": ["Title"],
 "preprocess": {
   "type":"TRIM",
   "meta":{"index": [18,-1]
   }
}
```

3. 爬取正文的规则

```
为减少规则撰写难度，正文规则只需要xpath即可
注意：由于我们资源有限，尽尽尽最大可能的少添加type为render的工作任务

如果是静态网页
{"type": "static","xpath": "//div[@class='mainContent pt0']","charset": "gbk"}
如果网页需要动态加载
{"type": "render","xpath": "//div[@class='mainContent pt0']"}
```