

爬虫基本规则整理

1. 静态网页

```
1 url: http://cqfb.people.com.cn/
2
3 regulation: {
4     "url_xpath": "//div[@class='news1']/h4/a",
5     "title_xpath": ""
6 }
```

8 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

```
1 普通的没有特殊字符串的：
2 比如返回格式为：
3 {
4     'msg': '',
5     'status': '',
6     'date': '',
7     'page_num': '',
8     'data': [
9         {
10             'title': '我爱中国',
11             'url': 'www.aizhongguo.com'
12         },
13         ''
14     ]
15 }
```

16 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：

```
17 {
18     'list_path': [0, data]
19     'url_path': ['data_field1', ..., url],
20     'title_path': ['data_field1', ..., title],
21     'preprocess': {'type': 'None'}}
22 }
```

23 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。

24
25
26 首尾有特殊字符串的：

```

27 比如返回格式为:
28  HSHFJKSHD data = (
29      {
30          'msg': '',
31          'status': '',
32          'date': '',
33          'page_num': '',
34          'data':[
35              {
36                  'title': '我爱中国',
37                  'url': 'www.aizhongguo.com'
38              },
39              ''
40          ]
41      }
42  )
43
44 规则为:
45  {
46      'list_path':[0,data]
47      'url_path':['data_field1',...,url],
48      'title_path':['data_field1',...,title],
49      'preprocess': {'type':'TRIM', 'meta':{'index': (18,-1)}}
50  }
51
52 示例网站:
53  http://qc.wa.news.cn/nodeart/list?
    nid=11100290&pgnum=1&cnt=50&attr=&tp=1&orderby=1&callback=jQuery1710003008
    7353680303686_1568706392177&_=1568706400699
54 对应示例规则:
55  {
56      "list_path":["data"],
57      "url_path": ["LinkUrl"],
58      "title_path": ["Title"],
59      "preprocess": {
60          "type":"TRIM",
61          "meta":{"index": [18,-1]
62      }
63  }

```

3. 爬取正文的规则

```

1  为减少规则撰写难度，正文规则只需要xpath即可
2  注意：由于我们资源有限，尽最大可能的少添加type为render的工作任务
3
4  如果是静态网页
5  {"type": "static","xpath": "//div[@class='mainContent pt0']","charset":
    "gbk"}
6  如果网页需要动态加载
7  {"type": "render","xpath": "//div[@class='mainContent pt0']}

```

