数据提取方法

json

一种数据交换格式，看起来像python类型（字典，列表）的字符串

- json.loads
  - 把json字符串转化成python类型
  - 用法：`json.loads(json字符串)`
- json.dumps（写入本地时候用）
  - 把python类型转化成json字符串
  - 用法：`json.dumps({"a":"b", "c":"d"})`
  - 在把数据写到本地的时候要用到两个参数
  - f.w rite(json.dumps({}, ensure_ascii = False, indent = 2))
  - ensure_ascii = False 表示不适用二进制编码的格式书写，可以显示中文
  - indent = 2 表示下一行在上一行基础上空两格，形式更美观

tips：url地址中含有`callback`字段，要删掉才能得到json字符串

xpath

- 首先要在安装`xpath helper`插件，下载了crx之后重命名成rar格式，解压之后在chrome浏览器扩展程序中加载到浏览器
- 使用快捷键：`shift+ctrl+x`
  - 单斜杠 `/` 从根目录向下寻找标签
  - 双斜杠 `//` 可以寻找任意标签
  - `[@ ]`可以选择某一属性对应的值，如`[//li@class='abcd']`选中的是li标签下class=abcd的内容
  - `@`放到方括号里面表示对某一节点进行修饰，放到外面表示取某一标签对应的值，如`/a/@b`表示取a标签里面b属性的值
  - 获取文本：使用`text()`，如`/a/text()`表示获取a标签下的文本，`/a//text()`获取a标签下所有的文本

lxml

第三方模块，使用前要安装

使用的是其中的`etree`类：

```
from lxml import etree
element = etree.HTML("html字符串")
element.xpath("上面xpath helper中的语法") # 获得需要的数据
```

写程序的时候，xpath路径要到网页源码中找，element中会出错

tips：一个处理字符串的方法

```
str.replace('a','b') # 将字符串中的a都替换成b
str.strip() # 去掉首位的空字符
str.strip("a") # 去掉首位的'a'字符
```

一些有用的处理方法

- 列表推导式，帮助快速生成列表
  - 如`[i+10 for i in range(10)]` // `[10,11,12, ...19]`
- 字典推导式，帮助快速生成字典
  - 如`{i: 1 for i in range(10)}` // `{10:1, 11:1, ...19:1}`
- 三元运算符，简化if语句
  - 如`a = 1 if b > c else 2` // 如果if后面表达式为真则将if前面的值传给等号前面的变量（a），如果为假则将else后面的值传给a

——by 秦小炅 2018.9.16