数据提取方法

json

- 数据交换格式,看起来像python类型(列表,字典)的字符串
- 使用json之前需要导入
- 哪里会返回json的数据
  - 流程器切换到手机版
  - 抓包app
- json.loads
  - 把json字符串转化为python类型
  - `json.loads(json字符串)`
- json.dumps
  - 把python类型转化为json字符串
  - json.dumps({})
  - json.dumps(ret1,ensure_ascii=False,indent=2)
    - ensure_ascii :让中文显示成中文
    - indent :能够让下一行在上一行的基础上空格
- 豆瓣电视爬虫案例

xpath和lxml

- xpath
  - 一门从html中提取数据的语言
- xpath语法
  - xpath helper插件:帮助我们从`elements`中定位数据
  - 1. 选择节点(标签)
    - `/html/head/meta` :能够选中html下的head下的所有的meta标签
  - 2. `//` :能够从任意节点开始选择

        - `//li` :当前页面上的所有的li标签
        - `/html/head//link` :head下的所有的link标签
    - 3. `@符号的用途`
        - 选择具体某个元素: `//div[@class='feed']/ul/li`
            - 选择class='feed'的div下的ul下的li
        - `a/@href` :选择a的href的值
    - 4. 获取文本:
        - `/a/text()` :获取a下的文本
        - `/a//text()` :获取a下的所有的文本
    - 5. 点前
        - `./a` 当前节点下的a标签
- lxml
    - 安装:pip install lxml
    - 使用

```
from lxml import etree
element = etree.HTML("html字符串")
element.xpath("")
```

基础知识点的学习

- 列表推导式
- 字典推导式
- 三元运算符

写爬虫的讨论

- 1. url
    - 知道url地址的规律和总得页码数:构造url地址的列表
    - start_url
- 2.发送请求,获取响应
    - requests

- 3.提取数据
  - 返回json字符串:json模块
  - 返回的是html字符串:lxml模块配合xpath提取数据
- 4.保存