简单爬虫系列之新闻爬取

经过上一章我们顺利拿到python下载地址之后，不管如何更新 我们都是可以获取到

并且大概了解urllib是如何运行，可以用一些爬虫框架shell去调试爬取的数据，这里不做讲解，一般用于scrapy shell来调试（用浏览器调试也可以）

> scrapy 是一个python第三方的一个爬虫框架，让爬虫更简单、需要自己去安装这个框架，才可以调用 不会安装的可以来QQ群：XXXXXXXX

- 目标URL：CCTV新闻
- 爬取数据：新闻标题、简介、和链接（可以自己去尝试爬取链接获取新闻内容，一定要自己动手尝试）

思路：GET请求CCTV新闻 获取数据 抓取

但实际上，网页源代码并没有新闻的这些内容，通过抓包分析到

http://news.cctv.com/data/index.json 这个`json` 是返回首页的新闻json数据

- 通过分析，这个json是所有新闻内容，页面只会加载一部分，当点击加载更多时候，再去加载json剩下的内容直到加载完出现

```
# 导入url请求模块（核心模块，不需要安装）
import urllib,urllib2,json
url = "http://news.cctv.com/data/index.json"

# GET请求新闻首页json数据
res = urllib.urlopen(url).read() # 返回对象.read() 读取 json返回文本

# 将字符串转为dict 字典
lis = json.loads(res)["rollData"] # loads注意s  返回rollData列表

# 迭代列表 遍历列表里面所有字典 通过输出，需要爬取写到文件可以用open("路径","w")对象
for i in lis:
    print u'ID：%s\n标题：%s\n概要：%s\nurl：%s\n时间：%s\n\n'%
(i["id"],i["title"],i["description"],i["url"],i["dateTime"])
```

获取到新闻ID，标题，概要，新闻url地址，和时间、

ID：ARTI6eYj7m0kwdueBmLpu5xy170111
标题：阿联酋5名外交官在坎大哈爆炸中遇难
概要：10号的爆炸发生在坎大哈省政府招待所内，伤者中包括坎大哈省省长胡、阿拉伯联合酋长国驻阿富汗大使朱马·卡比以及一名阿联酋使馆工作人员。阿联酋外交人员10号白天在坎大哈出席了一家阿联酋援建医院的启用仪式。
url地址：http://news.cctv.com/2017/01/11/ARTI6eYj7m0kwdueBmLpu5xy170111.shtml
时间：2017-01-11 16:17

ID：ARTImrKTtmInE8G6c8RT0624170111
标题：欧洲各界人士：期待习近平在达沃斯展现中国领导力
概要：中国国家主席习近平将首次出席世界经济论坛年会的消息公布后，受到欧洲各国人士广泛关注。多名权威人士接受中新社记者采访时表示，期待习近平在达沃斯展现中国领导力，为世界和平稳定、繁荣发展增添信心。

url地址：http://news.cctv.com/2017/01/11/ARTImrKTtmInE8G6c8RT0624170111.shtml
时间：2017-01-11 16:14

ID：ARTI77Dbp3fQd5kPbpfsuJXb170111
标题：外交部副部长:辽宁舰训练过程中往返台湾海峡是正常的
概要：外交部副部长刘振民今日表示，辽宁舰是中国的第一艘航空母舰，近几年一直在训练。台湾海峡是大陆与台湾共享的国际水道，所以辽宁舰训练过程中往返台湾海峡是正常的，对两岸关系不会有任何影响。
url地址：http://news.cctv.com/2017/01/11/ARTI77Dbp3fQd5kPbpfsuJXb170111.shtml
时间：2017-01-11 16:13

ID：ARTIXYuC29hplMrDxyDWii0r170111
标题：奥巴马告别演说重提“改变” 对美国前景更乐观
概要：当地时间10日，美国总统奥巴马回到他政治生涯的起点芝加哥，发表总统任期告别演说。他重提8年前竞选时的口号“改变”，呼吁美国人相信通过自身努力寻求“改变”的能力，强调他对美国的前景更加乐观。
url地址：http://news.cctv.com/2017/01/11/ARTIXYuC29hplMrDxyDWii0r170111.shtml
时间：2017-01-11 16:13

ID：ARTILtTOwAMJjmfjEWqKw7Q4170111
标题：山东青岛：女童险坠楼 邻里齐救援
概要：9日下午，青岛市黄岛区上演了惊险的一幕，一个五岁大的女孩在自家玩耍时，不小心从四楼阳台坠落，好在被三楼的晾衣架拦住，情况十分危急。
url地址：http://news.cctv.com/2017/01/11/ARTILtTOwAMJjmfjEWqKw7Q4170111.shtml
时间：2017-01-11 16:12

ID：ARTIHQDMFj7GGLD2jN2tWMGT170111
标题：外交部：中国亚太安全合作政策可用六个主题词阐释
概要：国务院新闻办公室11日发表《中国的亚太安全合作政策》白皮书并举行新闻发布会。外交部副部长刘振民指出，可以用合作共赢、开放创新、良性互动、对话协商、地区机制、务实合作六个主题词来阐释中国的亚太安全合作政策。
url地址：http://news.cctv.com/2017/01/11/ARTIHQDMFj7GGLD2jN2tWMGT170111.shtml
时间：2017-01-11 16:11

ID：ARTIYFjecjULiZ6M4VdpOGBb170111
标题：2/3受访者称奥巴马未能执行承诺
概要：美国当地时间10日晚，北京时间今天上午，即将卸任的美国总统奥巴马发表了他的“告别演讲”，宣布他八年总统生涯的结束。当天，一些到场的奥巴马支持者对于他任期内的工作予以了认可。但也有人对于近年来美国枪击频发，特别是芝加哥的高犯罪率问题表达了不满。
url地址：http://news.cctv.com/2017/01/11/ARTIYFjecjULiZ6M4VdpOGBb170111.shtml
时间：2017-01-11 16:10

ID：ARTI4bhMrK18eJZAxZhz4pwV170111
标题：美国贸易代表警告：特朗普贸易政策会削弱美竞争力
概要：美国贸易代表弗罗曼10号警告说，美国当选总统特朗普的贸易政策、特别是威胁对在美国境外生产但产品销回美国市场的企业征收关税的举措将削弱美国竞争力。
url地址：http://news.cctv.com/2017/01/11/ARTI4bhMrK18eJZAxZhz4pwV170111.shtml
时间：2017-01-11 16:09

ID：ARTIN07aUcjqKTyW6fV3O4BR170111
标题：美国：参议院密集审议新政府阁员提名
概要：当地时间本月20日中午，也就是北京时间21日凌晨，美国当选总统特朗普将宣誓就职。这也就意味着，特朗普政府将正式“接管”美国。
url地址：http://news.cctv.com/2017/01/11/ARTIN07aUcjqKTyW6fV3O4BR170111.shtml
时间：2017-01-11 16:08

ID：ARTI0YZuOI7kKpGeZ8hWBxr7170111
标题：海军：破冰船赴黄渤海 调查冰情防冰害
概要：近期受冷空气影响，渤海及黄海北部海域冰情有进一步加重趋势，水产养殖、海上航运、油气开发等海上生产受到严重影响。为实时获取海上冰情资料，做好冬季破冰减灾工作，日前，海军派出破冰船赴渤海和黄海北部海域开展冰情调查及救援任务。
url地址：http://news.cctv.com/2017/01/11/ARTI0YZuOI7kKpGeZ8hWBxr7170111.shtml
时间：2017-01-11 16:07

ID：ARTIFWreobJZqL0czoxsK4qu170111
标题：重庆推出首台（套）重大技术装备保险补偿试点
概要：为加快推进重大技术装备研制和应用，重庆日前出台《重庆市首台（套）重大技术装备保险补偿试点工作方案》，以降低创新成果应用风险，单张保单财政补贴最高可达100万元。
url地址：http://news.cctv.com/2017/01/11/ARTIFWreobJZqL0czoxsK4qu170111.shtml
时间：2017-01-11 16:06

ID：ARTID42pCb3fswq7sWSyLv8I170111
标题：阿联酋：5名外交官在坎大哈爆炸中遇难
概要：阿联酋政府11日证实，有5名外交官在10日发生在阿富汗坎大哈的爆炸袭击中遇难。
url地址：http://news.cctv.com/2017/01/11/ARTID42pCb3fswq7sWSyLv8I170111.shtml
时间：2017-01-11 16:06

ID：ARTILCRQtK7xSg9XaCZYmSfx170111
标题：海军：第二十四批护航编队访问沙特
概要：当地时间9日，刚刚完成亚丁湾、索马里海域护航任务的海军第二十四批护航编队抵达沙特阿拉伯吉达港，开始对沙特阿拉伯王国进行为期5天的友好访问。
url地址：http://news.cctv.com/2017/01/11/ARTILCRQtK7xSg9XaCZYmSfx170111.shtml
时间：2017-01-11 16:04

ID：ARTIb7tbNxUWBnqeQDCPEIwB170111
标题：网购七日无理由退货暂行办法今年3月15日起施行
概要：　　中新网1月11日电 据国家工商行政管理总局网站消息，6日，工商总局印发《网络购买商品七日无理由退货暂行办法 》。《办法》明确了不适用退货的商品范围和商品完好标准以及相关退货程序，并对网络商品销售者违反本办法规定，作出了明确的处罚细则。办法自2017年3月15日起施行。
url地址：http://news.cctv.com/2017/01/11/ARTIb7tbNxUWBnqeQDCPEIwB170111.shtml
时间：2017-01-11 16:02

> 以上是一些爬取数据。 分析得出：url尾部 和 id 判断相同
>
> 可以再次爬取新闻的详细内容：`urllib.urlopen( i["url"] ).read()`

> 抓包分析出还返回了这个json 热门搜索

通过这些可以爬取一些新闻数据，然后放置在自己的个人网页上，或者刚出的微信小程序里

遇到问题怎么办？赶紧加入python交流群：xxxxxxxxxxxxxxxxxx