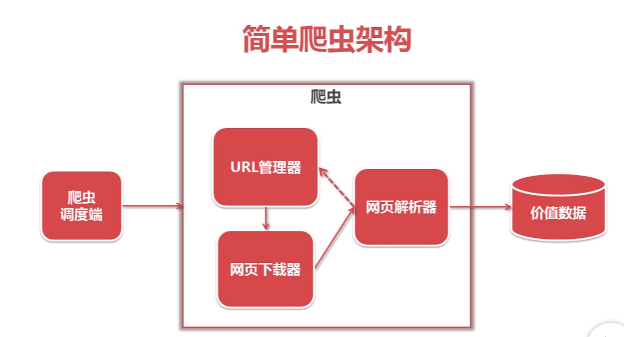
Python开发轻量级爬虫

不需要登录的静态网页的抓取

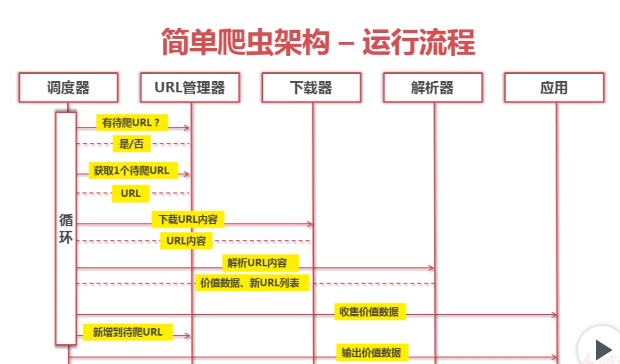
爬虫架构：URL管理器+网页下载器（urllib2）+网页解析器(BeautifulSoup)

爬虫：自动访问互联网并提取其中的价值数据

URL管理器：已经爬取得URL+待爬取的URL

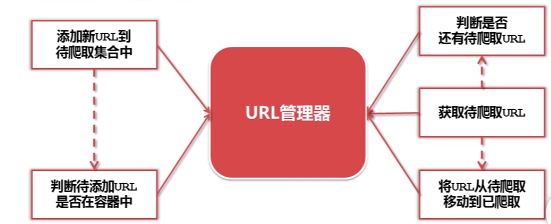


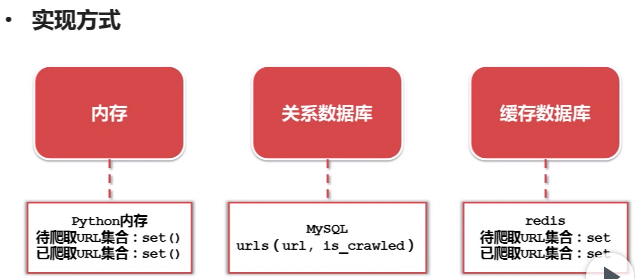
此简单爬虫结构的运行流程



URL管理器：管理待抓取URL集合和已抓取URL集合

-防止重复抓取，防止循环抓取





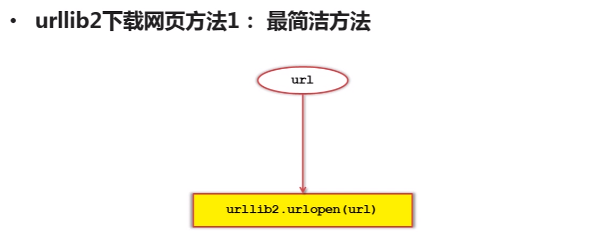
网页下载器：

将互联网上URL对应的网页下载到本地的工具，eg:**urllib2**

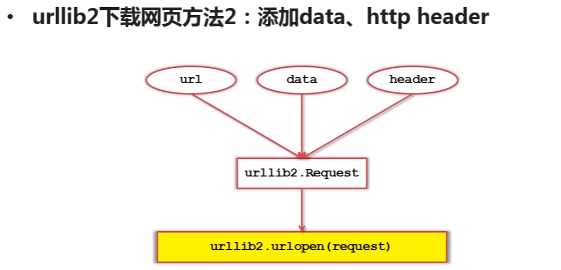
**Python3.x只有urllib urllib3**

**具体区别见https://yq.aliyun.com/articles/394693**

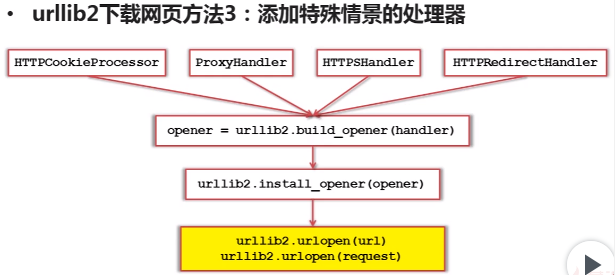




python2.X 中的 urllib.urlopen()被废弃, urllib2.urlopen()相当于python3.X中的urllib.request.urlopen()

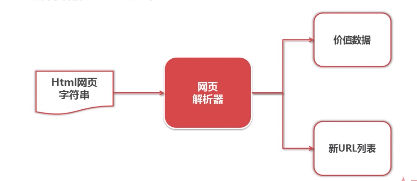


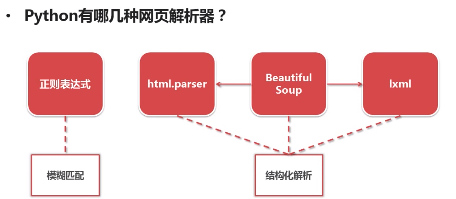




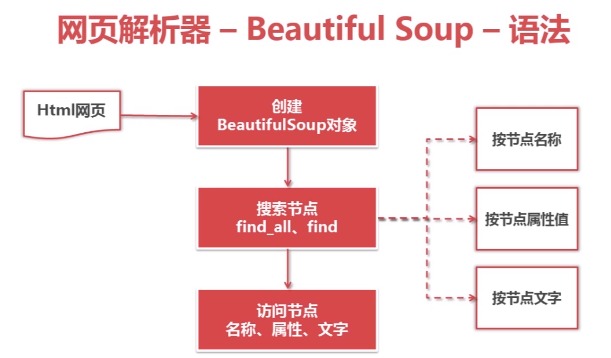


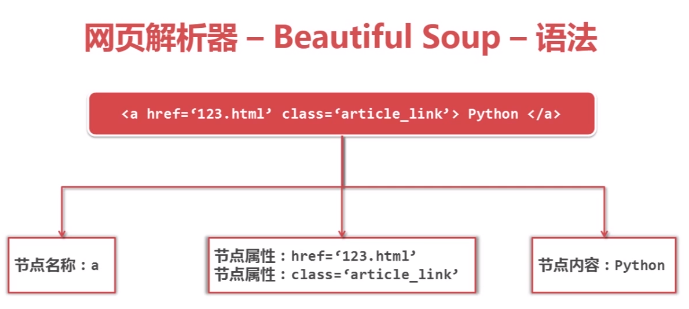
网页解析器：从网页中提取有价值数据的工具





结构化解析-DOM树

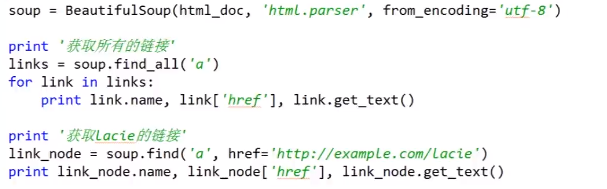


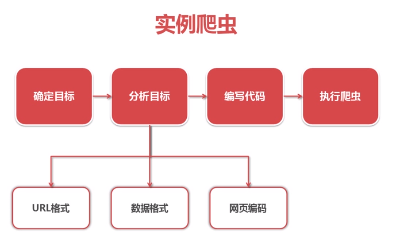


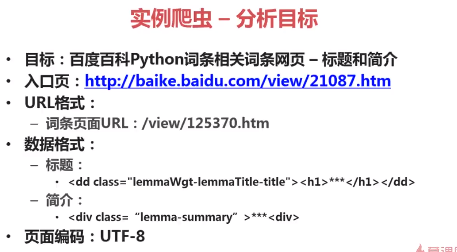












**URL**

<http://price.mofcom.gov.cn/pricequotation/priceSearchdetail.shtml?pro_name=&pro_trade=&pro_type=%E7%B2%AE%E5%86%9C%E5%9C%9F%E7%95%9C&pro_region=%E5%9B%BD%E5%86%85&startTime=&endTime>=

**数据格式**

-表格

-表项类名不变

<tr>

<th class="pl10 tal" width="20%">商品名称</th>

<th class="tac" width="10%">交易时间</th>

<th class="tac" width="30%">规格

<a href="pricequotationdetail.shtml?seqno=159">中国北方市场现货价</a>

//除row1外每行都有

</th>

<th class="tac" width="10%">单位名称</th>

<th class="tac" width="10%">价格</th>

</tr>

**页面编码**

UTF-8