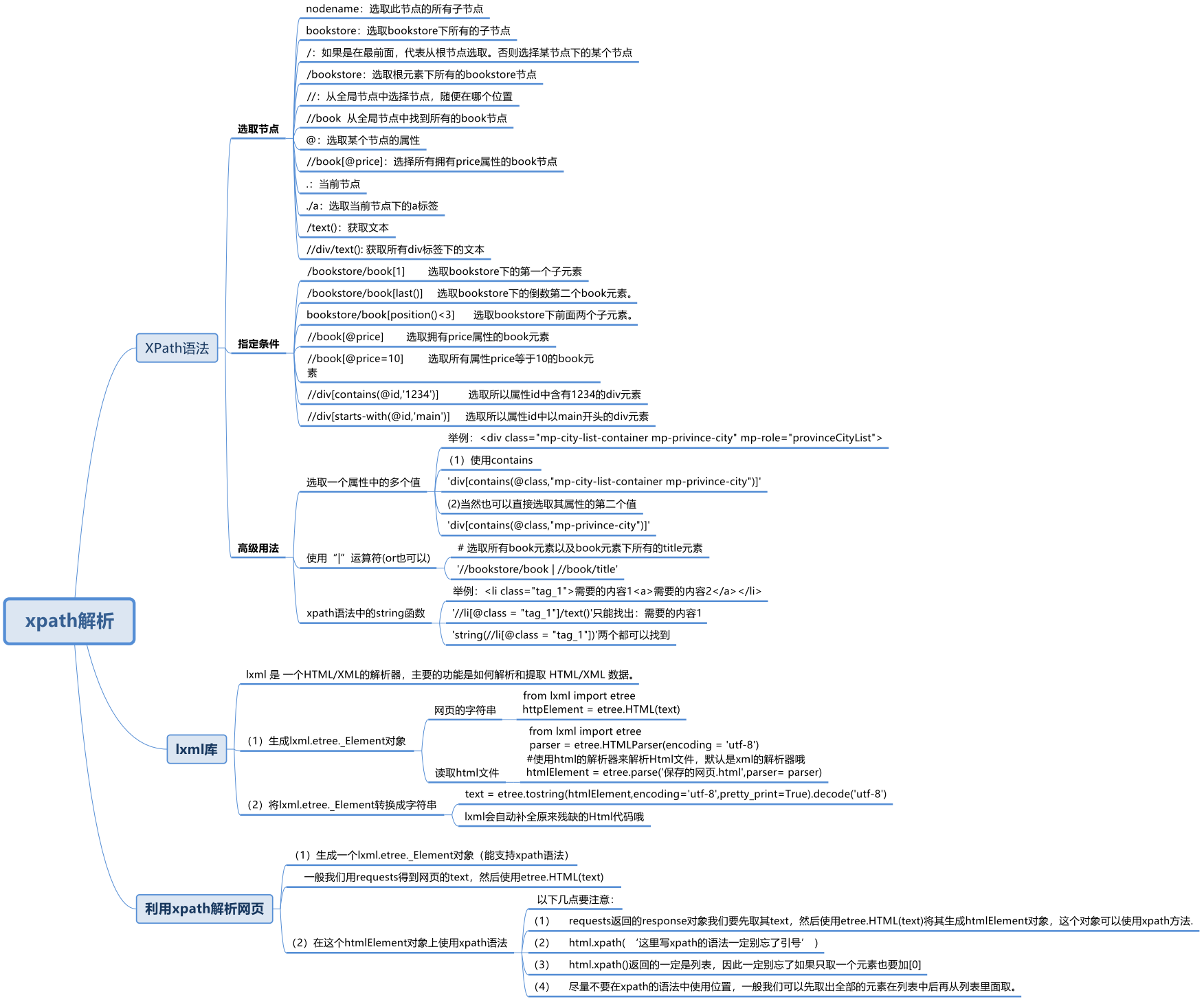
xpath解析
XPath语法
选取节点
nodename：选取此节点的所有子节点
bookstore：选取bookstore下所有的子节点
/：如果是在最前面，代表从根节点选取。否则选择某节点下的某个节点
/bookstore：选取根元素下所有的bookstore节点
//：从全局节点中选择节点，随便在哪个位置
//book 从全局节点中找到所有的book节点
@：选取某个节点的属性
//book[@price]：选择所有拥有price属性的book节点
.：当前节点
./a：选取当前节点下的a标签
/text()：获取文本
//div/text(): 获取所有div标签下的文本
指定条件
/bookstore/book[1] 选取bookstore下的第一个子元素
/bookstore/book[last()] 选取bookstore下的倒数第二个book元素。
bookstore/book[position()<3] 选取bookstore下前面两个子元素。
//book[@price] 选取拥有price属性的book元素
//book[@price=10] 选取所有属性price等于10的book元素
//div[contains(@id,'1234')] 选取所以属性id中含有1234的div元素
//div[starts-with(@id,'main')] 选取所以属性id中以main开头的div元素
高级用法
选取一个属性中的多个值
举例：
（1）使用contains
'div[contains(@class,"mp-city-list-container mp-privince-city")]'
(2)当然也可以直接选取其属性的第二个值
'div[contains(@class,"mp-privince-city")]'
使用“|”运算符(or也可以)
选取所有book元素以及book元素下所有的title元素
'//bookstore/book | //book/title'
xpath语法中的string函数
举例：<li class="tag_1">需要的内容1<a>需要的内容2
'//li[@class = "tag_1"]/text()'只能找出：需要的内容1
'string(//li[@class = "tag_1"])'两个都可以找到
lxml库
lxml 是 一个HTML/XML的解析器，主要的功能是如何解析和提取 HTML/XML 数据。
（1）生成lxml.etree._Element对象
网页的字符串
from lxml import etree
httpElement = etree.HTML(text)
读取html文件
from lxml import etree
parser = etree.HTMLParser(encoding = 'utf-8')
#使用html的解析器来解析Html文件，默认是xml的解析器哦
htmlElement = etree.parse('保存的网页.html',parser= parser)
（2）将lxml.etree._Element转换成字符串
text = etree.tostring(htmlElement,encoding='utf-8',pretty_print=True).decode('utf-8')
lxml会自动补全原来残缺的Html代码哦
利用xpath解析网页
（1）生成一个lxml.etree._Element对象（能支持xpath语法）
一般我们用requests得到网页的text，然后使用etree.HTML(text)
（2）在这个htmlElement对象上使用xpath语法
以下几点要注意：
（1） requests返回的response对象我们要先取其text，然后使用etree.HTML(text)将其生成htmlElement对象，这个对象可以使用xpath方法.
（2） html.xpath(‘这里写xpath的语法一定别忘了引号’)
（3） html.xpath()返回的一定是列表，因此一定别忘了如果只取一个元素也要加[0]
（4） 尽量不要在xpath的语法中使用位置，一般我们可以先取出全部的元素在列表中后再从列表里面取。

Xpath 扩展——XPath 轴

轴名称	结果
ancestor	选取当前节点的所有先辈（父、祖父等）。
ancestor-or-self	选取当前节点的所有先辈（父、祖父等）以及当前节点本身。
attribute	选取当前节点的所有属性。
child	选取当前节点的所有子元素。
descendant	选取当前节点的所有后代元素（子、孙等）。
descendant-or-self	选取当前节点的所有后代元素（子、孙等）以及当前节点本身。
following	选取文档中当前节点的结束标签之后的所有节点。
namespace	选取当前节点的所有命名空间节点。
parent	选取当前节点的父节点。
preceding	选取文档中当前节点的开始标签之前的所有节点。
preceding-sibling	选取当前节点之前的所有同级节点。
self	选取当前节点。

例子	结果
child::book	选取所有属于当前节点的子元素的 book 节点。
attribute::lang	选取当前节点的 lang 属性。
*child::**	选取当前节点的所有子元素。
*attribute::**	选取当前节点的所有属性。
child::text()	选取当前节点的所有文本子节点。
child::node()	选取当前节点的所有子节点。
descendant::book	选取当前节点的所有 book 后代。
ancestor::book	选择当前节点的所有 book 先辈。
ancestor-or-self::book	选取当前节点的所有 book 先辈以及当前节点（如果此节点是 book 节点）
child::/child::price*	选取当前节点的所有 price 孙节点。

Xpath 系统的学习资料：http://www.w3school.com.cn/xpath/index.asp