lxml

--class etree对象

```
from lxml import etree
```

--etree.HTML(网页源代码)

```
selector = etree.HTML(网页源代码)
```

--selector.xpath(xpath表达式)

进行提取

--表达式提取内容

//定位根节点

/往下层寻找

text()提取文本内容

FE：

- 不需要的信息1
- 不需要的信息2
- 不需要的信息3

```
//*[@id="useless"]/li[1]
selector = etree.HTML(html)
#print(selector.xpath('//*[@id="useless"]/li[1]/text()'))
for i in selector.xpath('//*[@id="useless"]/li/text()'):
      print(i)
```

或者：

```
for i in selector.xpath('//div/ul[@id="useless"]/li/text()'):
      print(i)
```

--表达式提取属性信息

/@属性名

FE：

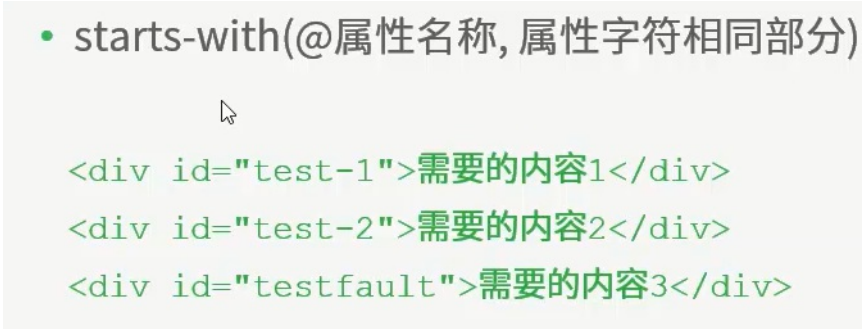
```
selector = etree.HTML(html)
link = selector.xpath('//a/@href')
for each in link:
      print(each)
```

结果：

```
>>>
================== RESTART: C:\Users\lenovo\Desktop\Tem.py ==================
http://jikexueyuan.com
http://jikexueyuan.com/course/
>>>
```

--starts-with(@属性名称,属性字符相同部分)

提取以相同的字符开头的



FE：

```
selector = etree.HTML(html1)
```

```
content = selector.xpath('//div[starts-with(@id,"test")]/text()')
for each in content:
      print(each)
```

--string(.)

标签套标签

FE:

```
print('='*8)
selector = etree.HTML(html2)
data = selector.xpath('//div[@id="test3"]')[0]
info = data.xpath('string(.)').replace('\n','').replace(' ','')
print(info)
```