https://www.runoob.com/xpath/xpath-tutorial.html

# XPath教程

XPath是一门在xml文档中查找信息的语言

XPath使用路径表达式在xml文档中进行导航

XQuery和XPointer均构建于XPath表达式之上

XPath包含一个标准函数库，有超过100个内建函数

这些函数用于字符串，数值，日期和时间比较，节点，QName处理，序列处理，逻辑值等。

# 读字符串: etree.HTML

## 例1

from lxml import etree

str='''<body>

<div>

<ul>

<li>01</li>

<li>02</li>

<li>03</li>

<li>04</li>

</ul>

</div>

</body>'''

html = etree.HTML(str)

li = html.xpath('//body/div/ul/li/text()')

for item in li:

print(item)

## 例2

from lxml import etree

str='''<p>

<ul>

<li class="item-1"><a href="1.html">item 1 </a></li>

<li class="item-2"><a href="2.html">item 2 </a></li>

<li class="item-3"><a href="3.html">item 3 </a></li>

<li class="item-4"><a href="4.html">item 4 </a></li>

<li class="item-5"><a href="5.html">item 5 </a></li>

</ul>

</p>'''

html = etree.HTML(str)

s=etree.tostring(html).decode()

print(s,"\n\n")

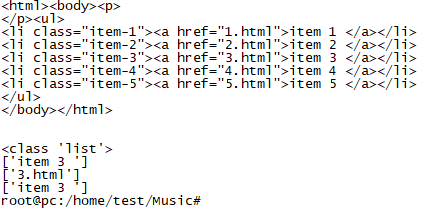
kk = html.xpath('//li[3]/a/text()')

print(type(kk))

print(kk)

print(html.xpath('//li[3]/a/@href'))

print(html.xpath('//li/a[@href="3.html"]/text()'))



# 读xml: read

from lxml import etree

with open('test.xml', 'r', encoding='utf-8') as f:

str = f.read()

#print(str)

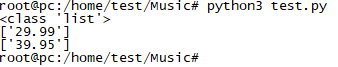
html = etree.HTML(str.encode('utf-8'))

kk=html.xpath('//bookstore/book[last()-1]/price/text()')

print(type(kk))

print(kk)

print(html.xpath('//bookstore/book[last()]/price/text()'))



# 读xml: etree.parse

## 例1

from lxml import etree

html=etree.parse('./test.xml', etree.HTMLParser())

res = etree.tostring(html)

print(res.decode('utf-8'))

print("\n\n")

#kk=html.xpath('//book')

#print(type(kk))

kk=html.xpath('//book/author/text()')

print(kk)

kk=html.xpath('//book/title/@lang')

print(kk)

root@pc:/home/test/Music# python3 test4.py

<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN" "http://www.w3.org/TR/REC-html40/loose.dtd">

<?xml version="1.0" encoding="UTF-8"??><html><body><bookstore>

<book>

<title lang="en">Harry Potter</title>

<author>J K. Rowling</author>

<year>2005</year>

<price>29.99</price>

</book>

</bookstore>

</body></html>

['J K. Rowling']

['en']

root@pc:/home/test/Music#

## 例2

<?xml version="1.0" encoding="UTF-8"?>

<bookstore>

<book>

<title lang="eng">Harry Potter</title>

<price>29.99</price>

</book>

<book>

<title lang="cn">Learning XML</title>

<price>39.95</price>

</book>

</bookstore>

from lxml import etree

html=etree.parse('./test.xml', etree.HTMLParser())

res = etree.tostring(html)

print(res.decode('utf-8'))

print("\n\n")

#kk=html.xpath('//book')

#print(type(kk))

kk=html.xpath('//book[2]/price/text()')

print(kk)

kk=html.xpath('//book[1]/price/text()')

print(kk)

kk=html.xpath('//book[last()]/price/text()')

print(kk)

kk=html.xpath('//book[position()>1]/price/text()')

print(kk)

print("\n\n")

kk=html.xpath('//title[@lang]/text()')

print(kk)

kk=html.xpath('//title[@lang="cn"]/text()')

print(kk)

print("\n\n")

kk=html.xpath('//book[price>35.00]/title/text()')

print(kk)

# 读requests

from lxml import etree

from lxml import html

import requests

#url="https://www.zhipin.com/c101280100/h\_101280100/"

url = "https://bj.58.com/ershoufang/p1/"

headers={'User-Agent':'Mozilla/5.0 (X11; Linux x86\_64; rv:98.0) Gecko/20100101 Firefox/98.0'}

#res=requests.get(url).text

res=requests.get(url, headers=headers, timeout=10).text

#print(res.encode('utf-8'))

#tree=etree.HTML(res)

tree=html.fromstring(res)

kk=tree.xpath('//span[@class="property-content-info-tag"]')

#kk=tree.xpath('//span[@class="property-content-info-tag"]/text()')

print(type(kk))

print(kk)

# 复杂应用：etree.parse



root@pc:/home/zcw# cat test.py

from lxml import etree

html=etree.parse('./test.html', etree.HTMLParser())

'''

kk=html.xpath('//div[@class="job-primary"]')

#print(len(kk))

#print(kk[0])

aa = kk[0].xpath('./div[@class="info-primary"]/div/@class')

#print(aa)

#aa1=kk[0].xpath('./div[@class="info-primary"]/div[@class="primary-wrapper"]')

aa2=kk[0].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div[@class="company-text"]/h3/@class')

#print(aa2)

aa2=kk[0].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div[@class="company-text"]/h3/a/@href')

aa2=kk[0].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div[@class="company-text"]/h3/a/@ititle')

aa2=kk[0].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div[@class="company-text"]/p/a/@href')

aa2=kk[0].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div[@class="company-text"]/p/a/@class')

#print(aa2)

#print(kk[0].xpath('./div[2]/@class'))

#print(kk[0].xpath('./div[1]/@class'))

#aa3=kk[0].xpath('./div[1]/div/div/p/a/@class')

#print(aa3)

job = kk[0].xpath('./div[@class="info-primary"]/div[@class="primary-wrapper"]/div[@class="primary-box"]')

#print(job[0].xpath('./div/@class'))

#print(job[0].xpath('./div[3]/@class'))

#print(job[0].xpath('./div[1]/span[1]/@class'))

title =job[0].xpath('./div[1]/span[@class="job-name"]/a/text()')

work\_addr=job[0].xpath('./div[1]/span[@class="job-area-wrapper"]/span/text()')

recruiter=job[0].xpath('./div[2]/div/h3/text()')

salary =job[0].xpath('./div[2]/span[1]/text()')

company =kk[0].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div/h3/a/text()')

company\_size=kk[0].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div/p/text()')

industry=kk[0].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div/p/a/text()')

url =kk[0].xpath('./div[@class="info-primary"]/div[@class="info-company"]/a/@href')

print(title[0])

print(work\_addr[0])

print(recruiter[0])

print(salary[0])

print(company[0])

print(company\_size[1])

print(industry[0])

print(url[0])

'''

kk=html.xpath('//div[@class="job-primary"]')

for i in range(len(kk)):

job = kk[i].xpath('./div[@class="info-primary"]/div[@class="primary-wrapper"]/div[@class="primary-box"]')

title =job[0].xpath('./div[1]/span[@class="job-name"]/a/text()')

work\_addr=job[0].xpath('./div[1]/span[@class="job-area-wrapper"]/span/text()')

recruiter=job[0].xpath('./div[2]/div/h3/text()')

salary =job[0].xpath('./div[2]/span[1]/text()')

company =kk[i].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div/h3/a/text()')

industry=kk[i].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div/p/a/text()')

company\_size=kk[0].xpath('./div[@class="info-primary"]/div[@class="info-company"]/div/p/text()')

url =kk[i].xpath('./div[@class="info-primary"]/div[@class="info-company"]/a/@href')

print(title[0])

print(work\_addr[0])

print(recruiter[0])

print(salary[0])

print(company[0])

print(company\_size[1])

print(industry[0])

print(url[0])

print("==============================================")

new\_links = html.xpath('//div[@class="page"]/a[@class="next"]/@href')

print(len(new\_links))

if new\_links and len(new\_links) > 0:

print(type(new\_links[0]))

print("https://www.zhipin.com"+new\_links[0])

root@pc:/home/zcw#

