Hints:

搜出前端后 ctrl+f 复制粘贴

剖析基本模版：

import requests

from bs4 import BeautifulSoup

newsurl='http://news.sina.com.cn/china/'

res=requests.get(newsurl)

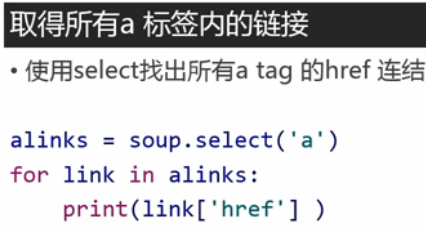
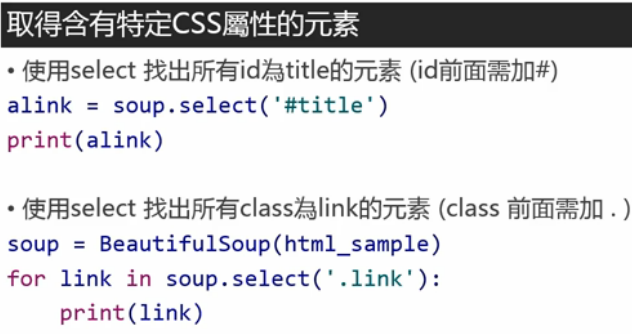
res.encoding='utf-8'

soup=BeautifulSoup(res.text,'html.parser')

print(soup.text)

BeautifulSoup:

.select(‘a’)形如此可以把前端指定标签里面元素取出来



对于右图，普适性来说，<>里面有什么属性我都可以把其中内容取出来，右图一般

注意！直接对初次soup对象进行select，其类型不是list所以可以直接跟href，但是对初次对象再select之后就是一个list了，此时取href需要注明第几个。

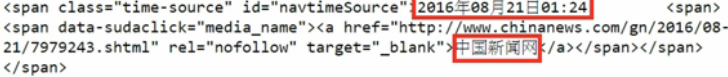
Chrome开发工具：在检查下刷新，观察network，信息一般存在doc里，左上角检查可以审查特定元素。他的class，css之类

在固定的模版中取出对象的各种消息：注意变量用soup赋值后，是soup类型，在之后的for循环中，循环变量也是soup类型，两者都可以.select如上图所讲有找tag，class和id这几种。

注意soup也是一个list，list没有text类型，只有形如x[0].text才可取出text。

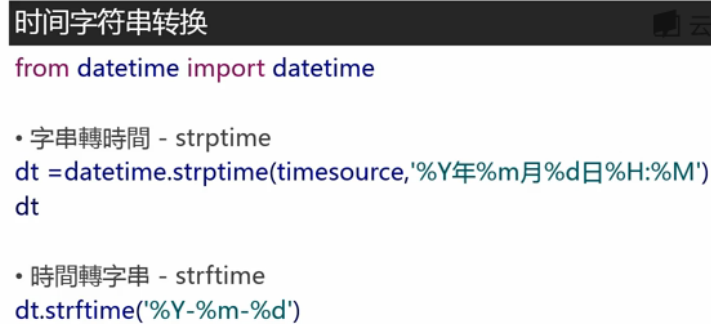
我们直接选出class下比如h2观察可能存在编码问题，但是取text编码是正常的。

Contents：



当select下存在多个text对象时，除了再细化搜出，我们还可以.contents[x]取出内容，若结尾出现无关的space或其他，用strip（）处理。

字符与时间转换：

注意2.7版本下，提取出的unicode一定要经过x.encode（‘utf-8’）编码后才可让strptime识别格式。

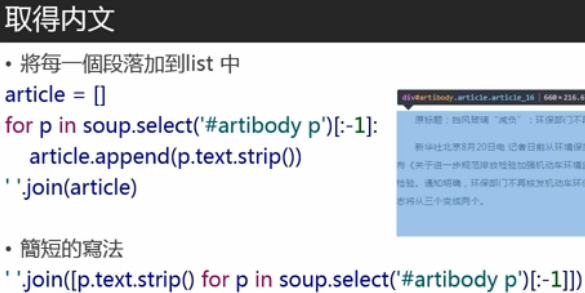
嵌套中提取：

Example



Timesource-span-a里的内容提取

多段同时提取：

这里在article id下有文章内容，分别放在三个p里，分别是标题、内容、编辑，[:-1]是倒数一个的意思相当于最后编辑信息不要，append的意思是对空list加内容，join的意思是以某种形式连接list中多个对象。

简单写法旨在直接对搜出来的p操作省去append过程。

strip使用技巧：strip是一个双向删除指定字符的函数，for example

authority=soup1.select('.show\_author')[0].text.lstrip(u'责任编辑:')

如此删除责任编辑，需要注意的是ascii遇到中文报错，一定要转utf-8

网址里提取编号：比如http://news.sina.com.cn/o/2018-01-22/doc-ifyquixe5830792.shtml

其中最后一个/中ify后.前部分为编号。

link1=link.split('/')[-1].lstrip('doc-ify').rstrip('.shtml')可如此分割/取最后一部分处理。

也可调用re模块，link2=re.search(r'doc-ify(.\*).shtml',link) print link2.group(1)

找资料方法：监听页面若发现有异步载入方式，不仅仅要在doc里找资料而且要在xhr和js里观察。

Format的用法：

