

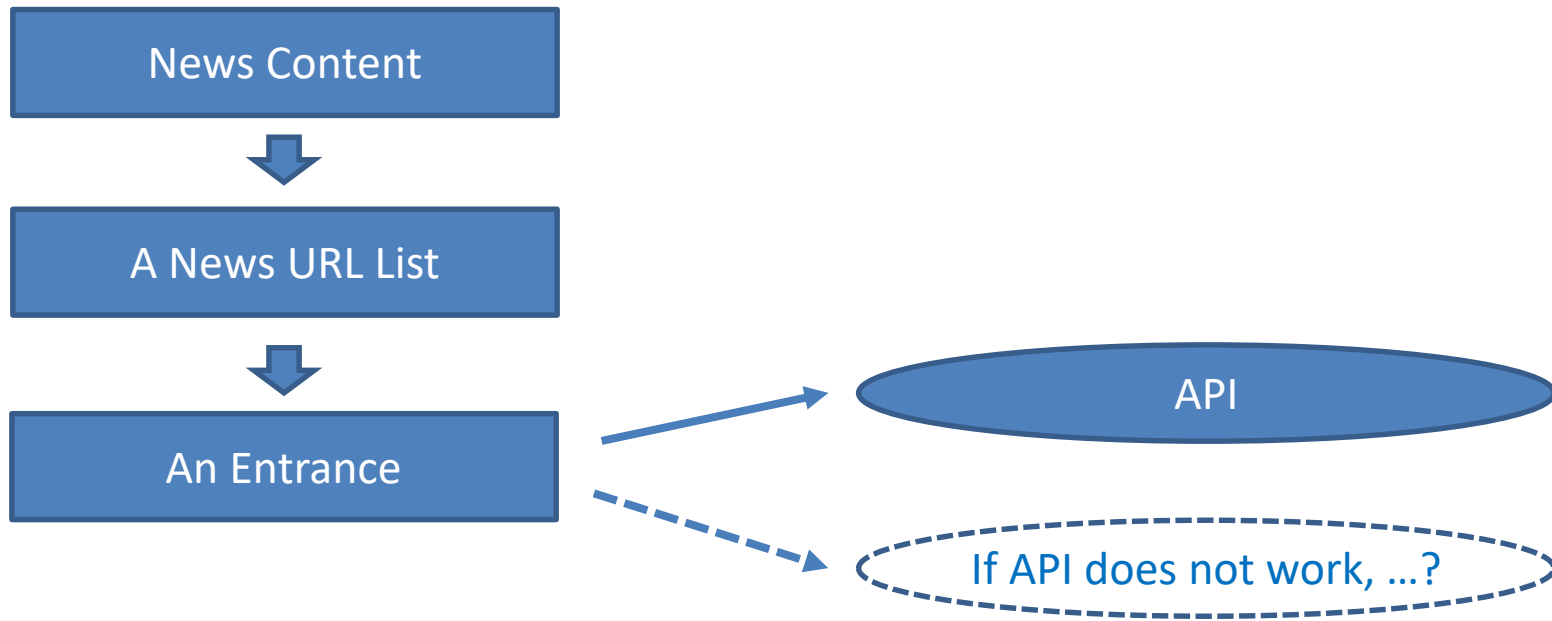


New Media Data Analytics and Application

Lecture 9: Crawlers for News Websites

Ting Wang

Break the Limitations of API



Directly start from website!

How to find the entrance?

The entrance should not be:

- Complex → Simple
- Unstructured → Structured

Data Structure is important!



Web Crawler

EXAMPLE 1:
Build a Web Crawler for
Eastday.com



URL List

- This is a root
- This is a start
- It has an entrance to the next page or the page of yesterday.

- 吴栏小学开展“扫黄打非”签名承诺活动 2018-12-12 14:54
- 金水区提升应急处置能力 上演水质“保卫战” 2018-12-12 14:54
- 2019英雄联盟春季赛即将开战 2018-12-12 14:54
- 不止更犀利，聊聊你不知道的斯巴鲁傲虎特装版 2018-12-12 14:54
- “笃笃笃”穿越千年而来的馄饨，承载着一代人的故乡情结 2018-12-12 14:54

第1页

[下一页](#)

[查看前一天](#)

Ⓜ 不安全 | mini.eastday.com/listpage/all.html

东方头条

做您关注的头条

婴儿坐车收费

小行星发现水迹象

碰瓷逼狗人

中关村融科起火

腾讯音乐IPO

搜索

头条 社会 军事 娱乐 国际 国内 健康 体育 图片 汽车 科技 财经 星座 更多

滚动新闻

全部

- 盛文兵：黄金区间震荡1241区域多，原油51.7区域参与多头 2018-12-12 19:29
- 短池世锦赛首日 中国队成绩喜人 2018-12-12 14:57
- 前三季度私募基金9656亿元投资于境内未上市未挂牌企业股权 2018-12-12 14:57
- 福建首部3D舞剧电影《丝路梦寻》在福州首映 2018-12-12 14:56
- 北演新年演出系列将再登大会堂 京剧名家集体亮嗓展风采 2018-12-12 14:56

- 别让童心过早被成人视角污染 2018-12-12 14:56
- “押宝”氢燃料动力 现代集团将投入67亿美元 2018-12-12 14:56
- 请给稚朴的纯真一份善意 2018-12-12 14:56
- 起诉寻人？是骚扰，不是痴情 2018-12-12 14:56
- 瓮福集团 主动开放 主动转型 2018-12-12 14:56

更多频道滚动新闻

国内 国际 社会 娱乐 时尚
军事 科技 汽车 体育 财经
健康 人文 游戏 星座 家居

全部

往期回顾

12月 2018年						
日	一	二	三	四	五	六
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

<http://mini.eastday.com/listpage/all.html>



上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

更多工具 开发者工具

浏览器地址栏: 不安全 | mini.eastday.com/listpage/all-20181211.html

页面内容:

- Road To F1: FIA单座方程式体系的通关秘籍 2018-12-11 23:48
- 刘德华演唱会即将开幕, 彭小龙再出新专辑 2018-12-11 23:48
- EXO都羡慕好奇自己为什么被叫“鹅王”? 遭队友KAI怼: 去照镜 2018-12-11 23:48
- 林忆莲新专辑走摇滚风 发片日撞蔡健雅艾怡良 2018-12-11 23:48
- NBA美女记者追随詹姆斯, 含情脉脉的眼神全是爱, 她是骑士的损 2018-12-11 23:48
- 曾与林忆莲齐名, 和张学友被称为金童玉女, 如今53岁依旧单身 2018-12-11 23:48
- 小龙女与新婚妻子坐在街头吃甜品略落魄, 成龙: 她开心就好 2018-12-11 23:48
- 千元性价比之王, 麒麟970+EMUI 9.0, 降至1499元 2018-12-11 23:48
- 取代微信不是梦想, 这个办公软件的发展, 给腾讯造成巨大的压力 2018-12-11 23:48
- 3换1? 莫雷敲定安东尼替代者, 火箭迎防守尖兵, 哈登: 他是核心 2018-12-11 23:48
- 文科生模拟考570分, 明年有希望进211大学吗? 这里有参照标 2018-12-11 23:48
- 喜报! 12月喜从天降, 外债还清, 4属相喜从天降, 生活蒸蒸日上 2018-12-11 23:48

第1页 下一页 查看前一天

浏览器开发者工具 (Elements):

```
auto;">...</ul>
<div class="pre_day_data clear-fix">
  ::before
  <div class="page_cnt clear-fix" style="display: block;">...</div>
  <a id="cat_yesterday" href="all-20181210.html" title="点击查看前一天">查看前一天</a> == $0
  ::after
</div>
</div>
<div class="run_right" id="right_fixed">...</div>
```

选中元素: a#cat_yesterday

Styles:

```
element.style {}
#cat_yesterday {
  float: right;
  color: #ee4b4b;
  font-size: 14px;
  text-decoration: underline;
  overflow: hidden;
  *zoom: 1;
  *display: inline-block;
```

Web Crawler

Eg.1 Solutions to Web Scraping Strategy of Eastday News



A Prototype of Web Crawler, Do you remember?

```
import urllib.request
response = urllib.request.urlopen('http://www.shisu.edu.cn/about/introducing-sisu')
HTMLText = response.read()

with open('Files/shisu.html', 'wb') as f:
    f.write(HTMLText)
```



• Step 1: re , Regular Expression 正则表达式

#找到入口, 获取URL List

```
import urllib.request
from bs4 import BeautifulSoup
import re
```

```
response = urllib.request.urlopen('http://mini.eastday.com/listpage/all-20181210.html')
HTMLText = response.read()
```

```
BSobj = BeautifulSoup(HTMLText, "html.parser") # 基于BeautifulSoup分析整个页面
```

```
ArticleLists = BSobj.find("ul", {"class": "item_run_news"})
```

```
for a in ArticleLists.findAll("a", href=True):#过滤出全部URL
    if re.findall('/a/', a['href']):#根据页面特征, 取得要抓取的URL
        print(a['href']) #超链接
        print(a.get_text()) #内容, 此处为新闻标题
```



- Step 2 : Discard Repeating URLs

```
#URL List去重, 结构化
import urllib.request
from bs4 import BeautifulSoup
import re

URLdict={}

response = urllib.request.urlopen('http://mini.eastday.com/listpage/all-20181210.html')
HTMLText = response.read()
BSobj = BeautifulSoup(HTMLText, "html.parser") # 基于BeautifulSoup分析整个页面

ArticleLists = BSobj.find("ul", {"class": "item_run_news"})

for a in ArticleLists.findAll("a", href=True):#过滤出全部URL
    if re.findall('/a/', a['href']):#根据页面特征, 取得要抓取的URL
        # print(a['href']) #超链接
        # print(a.get_text()) #内容, 此处为新闻标题
        URLdict[a['href']] = a.get_text()
print(URLdict)
```



Web Crawler

URL reform, as the case may be

mini.eastday.com/mobile/181209002140799.html

浏览器地址栏显示: mini.eastday.com/mobile/181209002140799.html

页面标题: 领导干部正确对待使用权力, 全心全意为人民服务

来源: 蜜蜂与花的故事

正文内容:

内江市资中县重龙镇谷田小学 张文盛

领导干部正确对待使用权力, 全心全意为人民服务

2018-12-09 00:21 蜜蜂与花的故事

领导干部如何对待和运用权力,是检验领导干部党性强弱、官德好坏的试金石。权力是把“双刃剑”,正确运用权力,掌权为公,执政为民,就会使权力造福人民;滥用权力,掌权为己,以权谋私,就会使权力造福人民。

邓小平同志曾经告诫说:“我们拿到这个权以后,就要谨慎。不要以为有了权就好办事,有了权就可以为所欲为,那样就非弄坏事情不可。”有一些领导干部在上任之初,尚能做到一身正气、两袖清风,但随着时间的延长,警惕性就不那么高了,就会在各种腐蚀面前败下阵来。因此,领导干部在运用权力时一定要以“如履薄冰、如临深渊”的谨慎行事。

要正确处理好权力与责任的关系。领导干部的级别越高,手中掌握的权力越大,社会责任就越重,越应更好地为人民服务。切不可滥用权力,谋取私利。领导干部必须认识权力是人民给的。我国宪法规定:中华人民共和国的一切权力属于人民。中国共产党作为执政党,是代表人民掌



领导干部正确对待使用权力, 全心全意为人民服务

2018-12-09 00:21 来源: 蜜蜂与花的故事

内江市资中县重龙镇谷田小学 张文盛

领导干部如何对待和运用权力,是检验领导干部党性强弱、官德好坏的试金石。权力是把“双刃剑”,正确运用权力,掌权为公,执政为民,就会使权力造福人民;滥用权力,掌权为己,以权谋私,就会使权力造福人民。

邓小平同志曾经告诫说:“我们拿到这个权以后,就要谨慎。不要以为有了权就好办事,有了权就可以为所欲为,那样就非弄坏事情不可。”有一些领导干部在上任之初,尚能做到一身正气、两袖清风,但随着时间的延长,警惕性就不那么高了,就会在各种腐蚀面前败下阵来。因此,领导干部在运用权力时一定要以“如履薄冰、如临深渊”的谨慎行事。

要正确处理好权力与责任的关系。领导干部的级别越高,手中掌握的权力越大,社会责任就越重,越应更好地为人民服务。切不可滥用权力,谋取私利。领导干部必须认识权力是人民给的。我国宪法规定:中华人民共和国的一切权力属于人民。中国共产党作为执政党,是代表人民掌

握领导权,每一位领导干部手中的权力都是人民赋予的。

党和人民血肉相连的关系,任何一名干部的成长都离不开党的培养和人民的信任,更离不开



Web Crawler

- Step 3: Put following code below the code of Step 2, to get the content of each news

```
for link, title in URLdict.items(): Reform: http://mini.eastday.com/a/181209002140799.html
    print(title, ":", link) #获得标题, 超链接, 保存
    #提取超链接的内容
    ContentResponse = urllib.request.urlopen('http://mini.eastday.com/mobile'+link[2:])
    ContentHTMLText = ContentResponse.read()
    ContentBSobj = BeautifulSoup(ContentHTMLText, "html.parser")
    ArticleSrcTime = ContentBSobj.find("div", {"class": "article-src-time"}) #获得时间、来源
    print(ArticleSrcTime.get_text()[0:17])
    print(ArticleSrcTime.get_text()[24:])
    Content=ContentBSobj.find("div", {"class": "J-article-content article-content"}) #获得详细内容, 保存
    print(Content.get_text())
```

Are you familiar
with these code?



```
for link, title in URLdict.items(): #提取新闻时间、来源、内容
```

```
try:
```

```
    ContentResponse = urllib.request.urlopen('http://mini.eastday.com/mobile'+link[2:]) #分析内容页
```

```
    ContentHTMLText = ContentResponse.read()
```

```
    ContentBSobj = BeautifulSoup(ContentHTMLText, "html.parser")
```

```
    ArticleSrcTime = ContentBSobj.find("div", {"class": "article-src-time"}) #获得时间、来源
```

```
    # print(ArticleSrcTime.get_text()[0:17]) #获得时间
```

```
    # print(ArticleSrcTime.get_text()[24:]) #获得来源
```

```
    Content=ContentBSobj.find("div", {"class": "J-article-content article-content"}) #获得详细内容, 保存
```

```
    NewsTitle=title.replace("\", \"' \").replace("\", \"' \")
```

```
    NewsLink='http://mini.eastday.com/mobile'+link[2:]
```

```
    NewsDate=ArticleSrcTime.get_text()[0:17]
```

```
    NewsAuthor=ArticleSrcTime.get_text()[24:]
```

```
    NewsContent=Content.get_text().replace("\", \"' \").replace("\", \"' \").replace(' ', '')
```

```
    # save data in database 连接到MySQL数据库
```

```
    # 1. Connection Open
```

```
    conn = pymysql.connect(user='root', password='123456', database='raw_data', charset='utf8')
```

```
    # 2. Cursor Creating:
```

```
    cursor = conn.cursor()
```

```
    # 3. SQL Execution# 执行SQL语句, 循环插入记录:
```

```
    sqlstr = "REPLACE INTO EASTDAY_ONLINE_NEWS(EASTDAY_NEWS_TITLE, EASTDAY_NEWS_DATE, EASTDAY_NEWS_AUTHOR, EASTDAY_NEWS_URL, EASTDAY_NEWS_CONTENT)
```

```
VALUES(' + NewsTitle + "',' + NewsDate + "',' + NewsAuthor + "',' + NewsLink + "',' + NewsContent + '')
```

```
    # 4. Cursor Moving# 执行, 游标移至当前位置
```

```
    cursor.execute(sqlstr)
```

```
    # 提交事务:
```

```
    conn.commit()
```

```
    # 5. Connection Close# 关闭Cursor:Connection:
```

```
    cursor.close()
```

```
    conn.close()
```

```
    SavedNum = 1
```

```
    print(NewsTitle + " " + NewsDate + " " + NewsLink)
```

```
except:
```

```
    print("Error:" + NewsTitle + " " + NewsLink)
```

```
finally:
```

```
    print(str(SavedNum) + " News Saved")
```

```
    SavedNum = 0 #SavedNum复位为0
```

Prepare for data base

```
# 顺藤摸瓜找到原文
```

```
import urllib.request
```

```
from bs4 import BeautifulSoup
```

```
import re
```

```
import pymysql
```

```
SavedNum = 0
```

```
URLdict= {}
```

```
response = urllib.request.urlopen('http://mini.eastday.com/listpage/all-20181210.html')
```

```
HTMLText = response.read()
```

```
BSobj = BeautifulSoup(HTMLText, "html.parser") # 基于BeautifulSoup分析整个页面
```

```
ArticleLists = BSobj.find("ul", {"class": "item_run_news"}) # 获取URL List
```

```
for a in ArticleLists.findAll("a", href=True): # 过滤出全部URL
```

```
    if re.findall('/a/', a['href']): # 根据页面特征, 取得要抓取的URL
```

```
        # print(a['href']) # 超链接
```

```
        # print(a.get_text()) # 内容, 此处为新闻标题
```

```
        URLdict[a['href']] = a.get_text()
```

```
# print(URLdict)
```



Ask A Question

How to get all data, not only one page?



#顺藤摸瓜找到原文

```
import urllib.request
from bs4 import BeautifulSoup
import re
import pymysql
```

SavedNum = 0

CurrentRootURL='http://mini.eastday.com/listpage/all-20181210.html' #常量转为变量, 为进入下一页做准备

```
while(True):
    URLdict = {}
    print("//////////////////// 开始--> "+CurrentRootURL)
    response = urllib.request.urlopen(CurrentRootURL)
    HTMLText = response.read()
    BSobj = BeautifulSoup(HTMLText, "html.parser") # 基于BeautifulSoup分析整个页面
```

Put these code at the
end of the program.



Step 5: To get the page link — use while()

```
NextPage = BSobj.find("div", {"class": "down_page"}) #获取下一页URL
if NextPage is None:
    #如果没有下一页, 则开始抓前一天的
    YesterdayPage = BSobj.find("a", {"id": "cat_yesterday"}) #获取前一天URL
    if re.findall('all-', YesterdayPage['href']): # 根据页面特征, 取得要抓取的URL
        YesterdayPageURL=YesterdayPage['href'] # 超链接
        CurrentRootURL = 'http://mini.eastday.com/listpage/' +YesterdayPageURL
else:
    #如果有下一页, 则抓下一页的
    for a in NextPage.findAll("a", href=True): # 根据页面特征, 取得要抓取的URL
        if re.findall('all-', a['href']):
            NextPageURL=a['href'] # 超链接
            CurrentRootURL = 'http://mini.eastday.com/listpage/' +NextPageURL
```



Web Crawler

Results:

Result Grid

Filter Rows:

Edit:

Export/Import:

Wrap Cell Content:

	EASTDAY_NEWS_ID	EASTDAY_NEWS_TITLE	EASTDAY_NEWS_DATE	EASTDAY_NEWS_AUTHOR	EASTDAY_NEWS_URL	EASTDAY_NEWS_CONTENT	EASTDAY_NEW
1		他用钢琴作为载体祝福广西壮族自治区繁荣...	2018-12-10 23:53:00	千龙网	http://mini.eastday.com/mobile/181210235300...	12月9日晚，庆祝广西壮族自治区成立60周...	NULL
2		苹果回应部分iPhone在中国遭禁售:是高通的...	2018-12-10 23:53:00	中国网	http://mini.eastday.com/mobile/181210235321...	10日晚间，针对高通称中国一法院同意其...	NULL
3		18支球队角逐第四届“大与磨力杯”羽毛球挑战...	2018-12-10 23:53:00	深圳新闻网	http://mini.eastday.com/mobile/181210235324...	12月9日，光明站的比赛在光明群众体育...	NULL
4		贝纳塞亚膝盖伤势缺席欧冠，鲁尼或获得...	2018-12-10 23:53:00	虎扑体育	http://mini.eastday.com/mobile/181210235317...	虎扑12月10日讯尤文图斯断俱乐部在官网...	NULL
5		曾经差点和赵又廷结婚，如今却依旧单身一人	2018-12-10 23:53:00	爱玩屎的阿拉蕾啊	http://mini.eastday.com/mobile/181210235318...	人这一生，找到能厮守终身的人其实并不...	NULL
6		李安祺中国“圈钱”，可拒绝用中文宣传，网...	2018-12-10 23:53:00	娱乐影狮说天下	http://mini.eastday.com/mobile/181210235321...	相信大家都是知道李小龙的，作为曾经的...	NULL
7		骚扰电话猖狂，直就治不了？	2018-12-10 23:52:00	扬子晚报网	http://mini.eastday.com/mobile/181210235222...	卖房子、卖保险、办贷款.....许多人被此类...	NULL
8		探访千年传统丁字 四川陵经砂器制作	2018-12-10 23:52:00	多彩贵州网	http://mini.eastday.com/mobile/181210235228...	12月10日，记者来到四川省荣经县探访...	NULL
9		《王者荣耀》即将加播曹老板，你对曹老板有...	2018-12-10 23:52:00	手机由音	http://mini.eastday.com/mobile/181210235202...	明天，《王者荣耀》将迎来一次更新，今...	NULL
10		魔兽世界！持续两个资料片的世界任务，你...	2018-12-10 23:52:00	暴雪游戏控	http://mini.eastday.com/mobile/181210235204...	自从《军团再临》资料片引入了“平行地图”...	NULL
11		王者荣耀：11月大更新，这些英雄将集体加...	2018-12-10 23:52:00	王者荣耀吧	http://mini.eastday.com/mobile/181210235305...	最近，王者荣耀的玩家们已经解开了大部...	NULL

online_news 1 x

Apply

Revert

Output

Action Output

#	Time	Action	Message	Duration / Fetch
1	18:35:07	SELECT * FROM raw_data.eastday_online_news limit 10000000	205471 row(s) returned	0.000 sec / 9.172 sec





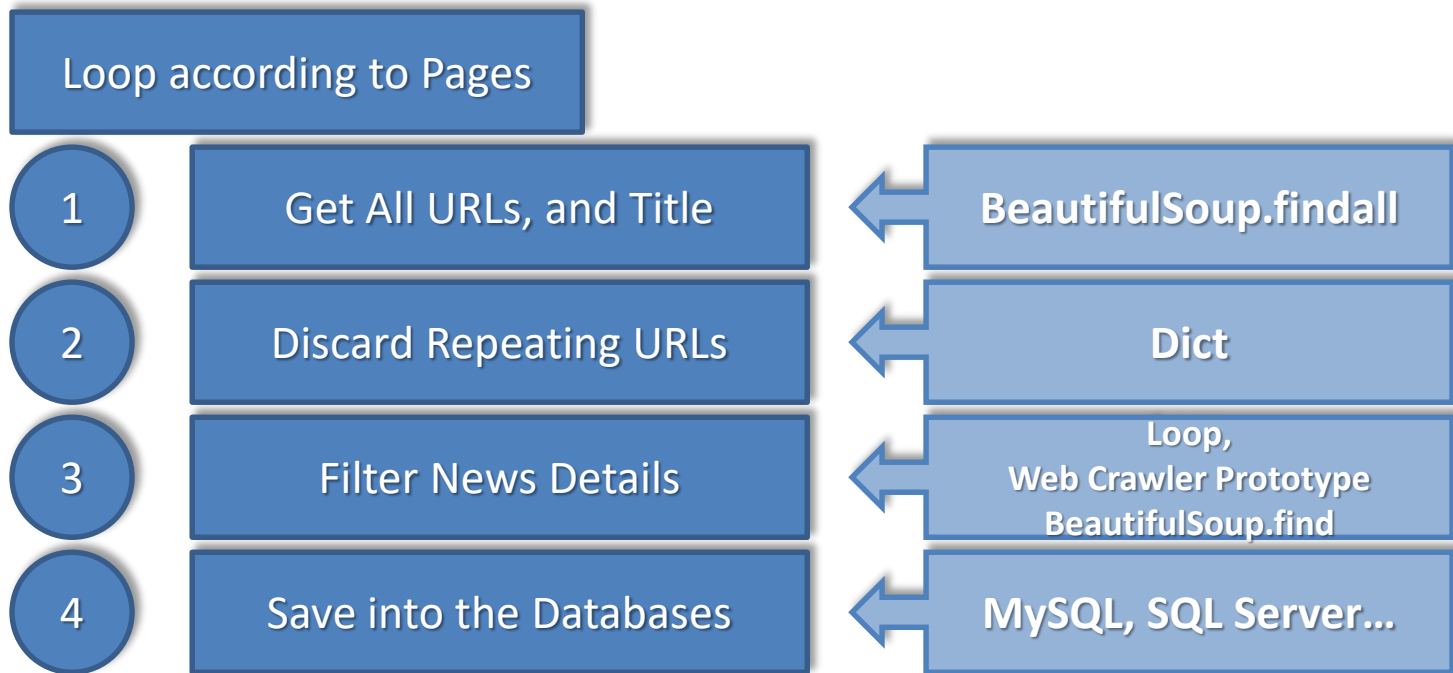
Ask A Question



Could you draw some conclusion for your Web Scraping Strategy of News List?

Web Crawler

- Steps:

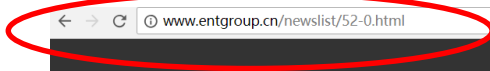


A photograph of a modern office interior. In the foreground, a white wall features a large, stylized purple logo on the left and the Chinese characters '芝恩' (Zhi'en) in black, with 'entgroup' in purple below them. The background shows a multi-level office space with glass railings, bright overhead lights, and some potted plants.

EXAMPLE 2:
Build a Web Crawler for
Entgroup

Web Crawler

<http://www.entgroup.cn/newslist/52-0.html>



We can find the new URLs,
if we click the menu.





Ask A Question



Why we can choose
<http://www.entgroup.cn/newslist/52-0.html>

*as a start page for web
scraping?*

- *It has a news list*
- *It has an entrance to the next page*



“IP买卖” 乱潮中 怎样才能让你的IP卖更多钱

根据不同IP形态进行有秩序、有逻辑的开发过程，这样的一个人可以称之为IP架构师或者IP经纪人。

2016-11-09 IP 电影产业 1



从《西部世界》回看今日，大数据会是下一个娱乐产业风口吗？

中国泛娱乐市场约有5000亿规模，成长空间巨大，而向成熟市场靠拢的娱乐工业化进程，数据驱动是必经路径。从行业专业分工和娱乐行业的特殊性的态势来

2016-11-08 资本 泛娱乐 大数据 1



情感品牌+超级明星，看娱乐营销如何打造商业超级IP
时下IP大热，纵观娱乐生态，内容和商业的边界非常模糊，但二者却是娱乐营销的重要方面，无论是内容还是品牌都在汲取IP红利。当品牌遇上IP，如何强势将娱
2016-11-17 营销 品牌 音乐 0



极客当道：化身“天使”守护原创者“顶层设计”孵化IP
投资市场、电影市场双双遇冷的当下，创业初期的“极客当道”就拿到了千万级的天使轮，是什么让投资方下重金布局的？
瀛文 2016-11-15 融资 影视 1



中国泛娱乐创新峰会”升级-未来电影“会场即将开启
在11月29-30日举办的2016中国泛娱乐创新峰会中，艺恩公司将举办“升级-未来电影”分会场，力邀电影产业各个环节的嘉宾共聚一堂，结合今年产业发展的新
2016-11-14 中国泛娱乐创新峰会 电影 1



精品网络剧强力反哺一线卫视 台网壁垒再被打破
2016年《老九门》、《九州天空城》、《如果蜗牛有爱情》三部网络剧相继反向输送一线卫视，在发行渠道多元化的同时，上量的网络剧通常在电视平台、网络
付晓岚 2016-11-14 自制剧 视频网站 3



发力直播+电商的背后，是一颗打造女性TVB帝国的雄心 | 对话兰渡CEO陆婷婷
虽然有时会嚷着“做内容太辛苦，下辈子我可能不会做这行了”，但陆婷婷还是带着她的兰渡文化在内容创业路上一路狂奔。
耿耀 2016-11-12 综艺节目 自制节目 3



三大视频网站付费会员实现4-5倍增长 自制内容成明年布局重点
继爱奇艺、腾讯视频宣布自家付费会员数量之后，日前，乐视视频也透露了相关信息。伴随着各家在内容投放数量上的加大，精品内容成为用户的主要诉求
瀛文 2016-11-11 视频网站 自制剧 自制节目 1



有钱，有项目，先导概念片你来拍！

更多工具 开发者工具

The screenshot shows the EntGroup website (www.entgroup.cn/news/Exclusive/) with the developer tools panel open. The website header includes the EntGroup logo, navigation links (分类, 艺恩汇, 中国票房), and user options (ENGLISH, 登录, 注册, 公司介绍). A search bar is present with the placeholder text "请输入您感兴趣的关键词". Below the search bar are tabs for "原创", "政策", "人物", "跨界", "资本", "营销", and "市场". The main content area displays three articles:

- EN Awards最佳娱乐营销案例优秀入围推荐 (一)**
2016中国泛娱乐指数盛典产业奖项为“最佳娱乐营销案例”，是EN Awards评比活动的常规保留奖项，历来备受业内关注。
2016-11-18 中国泛娱乐创新峰会 营销
- 重磅代表为项目站台 寻找未来影视新锐**
报名参加本次“时空引力场”的导演或编剧候选人将在活动后与片方面谈，并依据片方要求拍摄项目先导概念片，艺恩汇将全程支持“遇见未来-时空引力场”项目。
2016-11-18 中国泛娱乐创新峰会 影视
- 【艺恩观察】“轻科幻”成网剧新宠，“原创+IP”双轨制引行业新潮**
纵观近期上映的几部网剧，不难发现，以《如果蜗牛有爱情》《美人为馅》等为

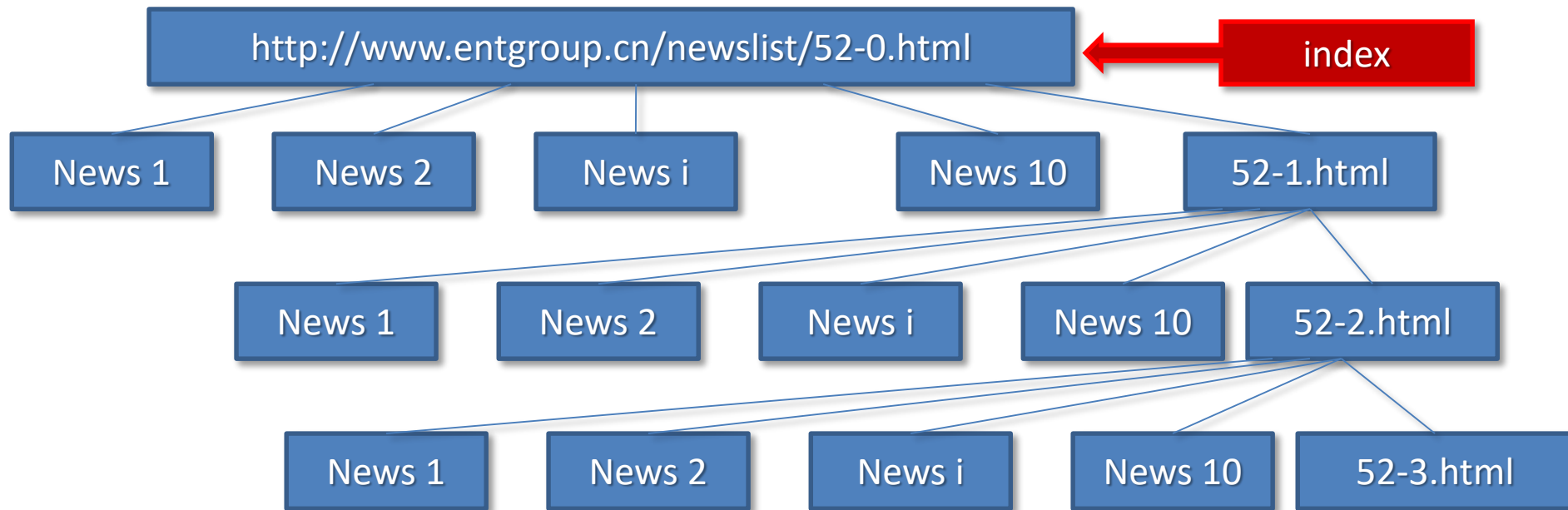
The developer tools panel on the right shows the HTML structure of the page. The selected element is the body, with the following CSS properties:

```
body {  
  background-color: transparent;  
  margin: 0px;  
  height: 100%;  
  display: block;  
  clear: both;  
  margin: 0px; width: 100%;  
}
```

The panel also shows the DOM tree and the computed styles for the selected element.

Web Crawler

Eg.2 Solutions to Web Scraping Strategy of EntGroup News



A Prototype of Web Crawler, Do you remember?

```
import urllib.request
response = urllib.request.urlopen('http://www.shisu.edu.cn/about/introducing-sisu')
HTMLText = response.read()

with open('Files/shisu.html', 'wb') as f:
    f.write(HTMLText)
```



- Web page analysis



Evolution based on the Prototype

```
import urllib.request
response = urllib.request.urlopen('http://www.entgroup.cn/newslist/76-0.html')
HTMLText = response.read()

with open('WebCrawler/entgroup/Markets/76.0.txt', 'wb') as f:
    f.write(HTMLText)
```

All the html codes have been downloaded.

We still need:

- *the URLs of news list*
- *the URL of next page*

How to get?



- Step 1 : Get All URLs, and Title

```
import urllib.request
from bs4 import BeautifulSoup
import re

response = urllib.request.urlopen('http://www.entgroup.cn/newslist/76-0.html')
HTMLText = response.read()

BSobj=BeautifulSoup(HTMLText,"html.parser")#基于BeautifulSoup分析整个页面
for a in BSobj.findAll("a", href=True):#过滤出全部URL
    if re.findall('/news/Markets/', a['href']):#根据页面特征，取得要抓取的URL
        print(a['href']) #超链接
        print(a.get_text()) #内容,此处为新闻标题

with open('WebCrawler/EntGroup/Markets/76.0.txt', 'wb') as f:
    f.write(HTMLText)
```



• Results of Step 1.

"C:\Program Files\Python36\python.exe" "D:/SISU/Courses/2018计算新闻学
/news/Markets/1358618.shtml

/news/Markets/1358618.shtml

2019或成中国科幻电影元年，作品质量有待观察

/news/Markets/1358617.shtml

/news/Markets/1358617.shtml

年末将至，来看看近期的13部献礼剧

/news/Markets/1258555.shtml

/news/Markets/1258555.shtml

近三分之一的直播平台倒闭，不少中小平台开始选择出海

/news/Markets/1258553.shtml

/news/Markets/1258553.shtml

文娱快讯|漫威将拍《奇异博士2》导演回归；赵薇&田壮壮合作首部话剧《求证》

/news/Markets/1258552.shtml

/news/Markets/1258552.shtml

《风味人间》看多了会饿 陈晓卿：那就对了

/news/Markets/1258551.shtml

/news/Markets/1258551.shtml

总结2018时，我们看到了些许中国动画电影的希望

www.entgroup.cn/newslist/76-0.html

艺恩网
entgroup.cn

分类

艺恩智库

中国票房

艺恩
entgroup

2018中国文娱创新峰会

2018年11月21日 北京· JW万豪酒店

原创

政策

人物

跨界

资本

营销

市场

电影

视频



2019或成中国科幻电影元年，作品质量有待观察

“中国科幻电影想要有突破，就得先往前走，不然一点机会都没有。”

2018-12-13 科幻 电影 国产片

0



年末将至，来看看近期的13部献礼剧

年末献礼剧多以现实题材为主，故事更具真实性，人物更加立体，也赋予了更多的艺术思考与人文关怀

2018-12-13 电视剧 现实题材

0



近三分之一的直播平台倒闭，不少中小平台开始选择出海

中小平台陆续出海，直播下半场将游进的是红海还是蓝海？



- Step 2 : Discard Repeating URLs

```
import urllib.request
from bs4 import BeautifulSoup
import re
URLdict=dict()#建立dict,分配空间
response = urllib.request.urlopen('http://www.entgroup.cn/newslist/76-0.html')
HTMLText = response.read()
BSobj=BeautifulSoup(HTMLText,"html.parser")#基于BeautifulSoup分析整个页面
for a in BSobj.findAll("a", href=True):#过滤出全部URL
    if re.findall('/news/Markets/', a['href']):#根据页面特征, 取得要抓取的URL
        #print(a['href']) #超链接
        #print(a.get_text()) #内容,此处为新闻标题
        URLdict[a['href']]=a.get_text() #装入dict, 以超链接为唯一key, 去除重复的
with open('WebCrawler/EntGroup/Markets/76.0.txt', 'wb') as f:
    f.write(HTMLText)
print(URLdict)
```



• Results of Step 2.

04EntgroupExclusiveMainWebCrawlerStep2



```
"C:\Program Files\Python36\python.exe" "D:/SISU/Courses/2018计算新闻学(研)/Lecture 6/04EntgroupExclusiveMainWebCrawlerStep2.py"
```

```
(' /news/Markets/1358618.shtml': '2019或成中国科幻电影元年, 作品质量有待观察', ' /news/Markets/1358617.shtml': '年末将至, 来看看近期的13部献礼剧', ' /news/Markets/1258555.shtml': '近三分之一的直播平台倒闭, 不少中小平台开始选择出海',  
 ' /news/Markets/1258553.shtml': '文娱快讯|漫威将拍《奇异博士2》导演回归; 赵薇&田壮壮合作首部话剧《求证》', ' /news/Markets/1258552.shtml': '《风味人间》看多了会饿 陈晓卿: 那就对了', ' /news/Markets/1258551.shtml': '总结2018时, 我们看到了些许中国动画电影的希望', ' /news/Markets/1258537.shtml': '星美回应倒闭传闻 公司近期将回归正轨', ' /news/Markets/1258535.shtml': '在互联网的推动下, 剧集似乎迎来了最好的时代', ' /news/Markets/1258533.shtml': '第49周剧集榜:《你和我的倾城时光》播放指数75.7 剧集榜头名,《将夜》领跑网络剧榜', ' /news/Markets/1258531.shtml': '为何Netflix原创内容仍不如版权内容受欢迎?', ' /news/Markets/1258530.shtml': '唯流量论时代将落幕, 会员左右视频网站竞争', ' /news/Markets/1258505.shtml': '陌陌研发力度加码 不久或将“放大招”', ' /news/Markets/1158503.shtml': '文娱快讯: 于正新剧《盛唐攻略》启动; 迪士尼2018年度票房突破70亿美元', ' /news/Markets/1158501.shtml': '樊路远: 阿里对大文娱、优酷、内容产业投入的决心、信心和耐心都不会改变', ' /news/Markets/1158451.shtml': '2018文艺片市场观察: 7亿和7万票房同在, 高口碑和低品质共存', ' /news/Markets/1158450.shtml': '在游戏短视频行业中应运而生的MCN机构', ' /news/Markets/1158449.shtml': '第49周电影榜: 海王6亿强势拉升票房 其余电影生存空间受挤压', ' /news/Markets/1158447.shtml': '斗鱼上天, 但直播行业未来难料', ' /news/Markets/1158441.shtml': '2018年风口大盘点: 短视频社交新起, 知识付费新电商破局', ' /news/Markets/1158438.shtml': '资本寒冬, 中国流媒音乐台挺直了腰板: “我们不一样!”', ' /news/Markets/1058398.shtml': '文娱快讯: 绫野刚搭档宋佳主演《破阵子》; 吴京否认《战狼3》被叫停', ' /news/Markets/0858347.shtml': '文娱快讯:《海王》首日票房超1.5亿 第61届格莱美音乐奖完整提名公布', ' /news/Markets/0958349.shtml': '《狗十三》2天仅1600万, 说好的小众电影春天呢?', ' /news/Markets/0958351.shtml': '星美被剔除恒生指数, 影院行业面临洗牌')
```

Process finished with `exit` code 0



- Step 3 : Get News Details

```
import urllib.request
from bs4 import BeautifulSoup
import re
URLdict=dict()#建立dict,分配空间
response = urllib.request.urlopen('http://www.entgroup.cn/newslist/76-0.html')
HTMLText = response.read()
BSobj=BeautifulSoup(HTMLText,"html.parser")#基于BeautifulSoup分析整个页面
for a in BSobj.findAll("a", href=True):#过滤出全部URL
    if re.findall('/news/Markets/', a['href']):#根据页面特征,取得要抓取的URL
        #print(a['href']) #超链接
        #print(a.get_text()) #内容,此处为新闻标题
        URLdict[a['href']]=a.get_text() #装入dict,以超链接为唯一key,去除重复的
for link,title in URLdict.items():
    print(title,":", link) #获得标题,超链接,保存
    #提取超链接的内容
    ContentResponse = urllib.request.urlopen('http://www.entgroup.cn'+link)
    ContentHTMLText = ContentResponse.read()
    ContentBSobj = BeautifulSoup(ContentHTMLText, "html.parser")
    Content=ContentBSobj.find("div",{"class":"detailsbox"})
    #获得详细内容,保存
    print(Content.get_text())
with open('WebCrawler/EntGroup/Markets/76.0.txt', 'wb') as f:
    f.write(HTMLText)
```



• Results of Step 3.

2019或成中国科幻电影元年，作品质量有待观察：</news/Markets/1358618.shtml>

2019或成中国科幻电影元年，作品质量有待观察

标签：

科幻电影国产片

来源：壹娱观察作者：2018-12-13

艺恩网转载本文只以信息传播为目的,不代表认同其观点和立场

[摘要]

“中国科幻电影想要有突破，就得先往前走，不然一点机会都没有。”

“离死只有一线之隔”。从十几米的空中落下来，防护服透明头盔面罩上白茫茫一片，看不清吴孟达的脸，他的呼吸很弱了，花了40秒钟，电钻取开封闭面罩上的螺丝后，吴孟达的脸和头发一样白，工作人员急忙把吴孟达送到片场吸氧。66岁的吴孟达此前做了心脏手术，但他坚持不用替身亲自上阵吊威亚。虽然防护服里有供气系统，但吴孟达拍戏吊威亚这件事已经多年没做过了，何况还要穿着80斤重的戏服。为此导演郭帆从一开始就要求剧组人员里有一名全科医生。“老爷子回到宾馆，第一件事就是哭：我60多岁了，为什么要干这件事？”回忆起《流浪地球》拍摄时的这一幕，制片人龚格尔扶了下眼镜，用手抹掉眼泪。台下暗自垂泪，台上谈笑风生。《流浪地球》定档发布会上，这个生死瞬间被吴孟达一笑而过。“我不好意思拒绝导演”，吴孟达幽默的解释引来了台下的欢笑，“有生之年参加中国自己的科幻电影，义不容辞，也此生无憾”。郭帆和龚格尔在写《流浪地球》的剧本时，吴孟达的名字已经浮现在他们的脑海中。尤其是龚格尔对于吴孟达的推崇更是堪称“粉丝”。如果你看过胡戈的《007大战猪肉王子》等一系列恶搞短片，你不会对他感到陌生，也更能理解这个高歌“被逼的”的龚格尔，对喜剧演员吴孟达的赞许。坐在郭帆导演工作室里的龚格尔与十年前的镜头里的“猪肉王子”判若两人，甚至跟三年前刚加入《流浪地球》剧组时的青涩都迥然相异。除了为吴孟达动容，现在你看不到他太多的表情变化，声音里透出的是冷静和沉着。手里的烟熟练点起，这是筹备《流浪地球》三年来他最大的表象改变。但你还是能看出他眼神里的亮色，他向我们描述着《流浪地球》的观感：“一开始是兴奋之中带着悲壮感，之后是入戏阶段的眼前一亮，再往下是……最终一路激昂直奔云霄，你出了电影院一定大汗淋漓！”。这个稍微夸张的观感并非无的放矢，根据数据显示：轻度观众、普通观众、中重度科幻观众普遍对《流浪地球》的满意度评分不低，而最近《流浪地球》放出的预告片显示其视效品质不会成为影片短板。人们对2019年满怀期待。中国电影院线将迎来首部真正的国产科幻大片并打开类型市场。除了2018年年初上映的《流浪地球》、《巨齿鲨2》、《上海堡垒》、《坏蛋联盟》等几部国产科幻电影外，在2018年龚格尔再次临阵，2018年是“中国科幻电影元年”，但其电影《三体》亦



- Step 4 :

```
import urllib.request
from bs4 import BeautifulSoup
import re

URLdict=dict()# 建立dict, 分配空间

for i in range(10):
    response = urllib.request.urlopen('http://www.entgroup.cn/newslist/76-'+str(i)+' .html')
    HTMLText = response.read()
    BSobj=BeautifulSoup(HTMLText,"html.parser")# 基于BeautifulSoup 分析整个页面
    for a in BSobj.findAll("a", href=True):# 过滤出全部URL
        if re.findall('/news/Markets/', a['href']):# 根据页面特征, 取得要抓取的URL
            #print(a['href']) #超链接
            #print(a.get_text()) #内容,此处为新闻标题
            URLdict[a['href']]=a.get_text() #装入dict, 以超链接为唯一-key, 去除重复的
    for link,title in URLdict.items():
        print(title, ":", link) #获得标题, 超链接, 保存
        #提取超链接的内容
        ContentResponse = urllib.request.urlopen('http://www.entgroup.cn'+link)
        ContentHTMLText = ContentResponse.read()
        ContentBSobj = BeautifulSoup(ContentHTMLText, "html.parser")
        Content=ContentBSobj.find("div",{ "class": "detailsbox" })
        #获得详细内容, 保存
        print(Content.get_text())
    with open("WebCrawler/EntGroup/Markets/76-'+str(i)+' .txt', 'wb') as f:
        f.write(HTMLText)
```



• Results of Step 4.

Run: 06EntgroupExclusiveMainWebCrawlerStep4 x

“C:\Program Files\Python36\python.exe” “D:\SISU\Courses\21
百度音乐: 起个大早赶个晚集 : /news/Markets/1358623.shtml

百度音乐: 起个大早赶个晚集

标签:

来源: 音乐先声作者: 麻炉2018-12-13

艺恩网转载本文只以信息传播为目的, 不代表认同其观点和立场
[摘要]

作为太合音乐的大股东, 百度的这一举动难免耐人寻味

名称	修改日期	类型	大小
76.0.txt	2018/12/13 11:14	文本文档	36 KB
76.1.txt	2018/12/13 11:14	文本文档	36 KB
76.2.txt	2018/12/13 11:14	文本文档	38 KB
76.3.txt	2018/12/13 11:14	文本文档	37 KB
76.4.txt	2018/12/13 11:15	文本文档	36 KB

??如今说起百度音乐, 仿佛是在说上世纪的老古董, 这一个如今已经无迹可寻的音乐产品, 却承载了MP3时代音乐的辉煌。只可惜, 在时代的米依傍阿里爸爸的当下, 百度对音乐业务的掌控早已脱了节。尽管已被市场宣布提前离场, 但音乐先声也注意到, 百度也在努力通过投资或是硬件合作的形式, 试图分得音乐产: 发展了十多年, 网络音乐市场刚刚萌芽。QQ音乐、酷我音乐、酷狗音乐都刚成立不久, 百度也颇有远见地瞄准了这个市场, 果断收购了当时火热的千千静听播放器, 正式进军数: 载歌曲, 然后用千千静听来循环播放。在当时娱乐形式还很有限的时代, PC基本只能用来打打游戏听听歌, 因此占据了听歌入口的百度也走上了发展的快车道。据公开资料, 早: 玩好。在收购千千静听之后, 百度音乐播放器与千千静听播放器一起共存了数年时间, 两者功能冲突, 百度音乐播放器的地位也略显尴尬。2011年, 百度又推出了旗下音乐平台: 识。直到2012年, 百度才将旗下所有产品整合更名为百度音乐。不过此时, 百度已然错过了打响品牌最关键的那几年。对外部环境的变化没有及时响应, 也是百度的失误之一。在互联网带宽有限的时代, 下载音乐再用播放器来播放是常见的使用场景, 但随着互联网带宽的迅速发展和移动智能设备的普及, 在线播放逐渐兴起, 下载已经成了不必要的动: 会歌曲模式, 甚至推出了歌单等个性化推荐的功能, 符合了用户的需求。在这种情况下, 下载 mp3 单纯的播放器, 都成了行将这个时代淘汰的物件, 而移动互联网时代的到来, 也



Ask A Question

How to save all these data?

Data Base



Prepare your Databases

档案名称	FILM_NEWS					
档案用途	影视新闻资料档					
主键(PK)	FILM_NEWS_PK: NEWS_ID(Cluster Index)					
附键(AK)	FILM_NEWS_AK: ORIGINAL_URL(Cluster Index)					
INDEX NAME		栏位	用途			
序号	栏位名称	栏位说明	资料形态	长度	Null	Default
01	NEWS_ID	新闻编号	Number		X	
02	NEWS_TITLE	新闻标题	Varchar		X	
03	NEWS_CONTENT	新闻内容	Text		X	
04	ORIGINAL_URL	原始网址	Varchar		X	
05	Publish_Date	新闻时间	Date			

SQL Script

```
CREATE TABLE FILM_NEWS
(
    NEWS_ID          INT(20)    PRIMARY KEY AUTO_INCREMENT,
    NEWS_TITLE       Varchar(500),
    NEWS_CONTENT     TEXT,
    ORIGINAL_URL     Varchar(500) NOT NULL UNIQUE,
    PUBLISH_DATE     DATE
);
```



Store into Databases

1. Discard Repetitions

- Dict()
- ORIGINAL_URL NOT NULL UNIQUE,

Try:

Insert into database

Except:

Not unique mistakes / other mistakes

Finally:

Next item



Store into Databases

2. Insert into Databases

```
sqlstr = "INSERT INTO FILM_NEWS(NEWS_TITLE,NEWS_CONTENT,ORIGINAL_URL,PUBLISH_DATE)  
VALUES('"+title+"','"+Content.get_text()+"','http://www.entgroup.cn"+link+"','"+PublishDate.get_text()+"');"  
  
cursor.execute(sqlstr)  
  
conn.commit()
```

Finish it by yourself!



Web Crawler

Results



Limit to 1000 rows

1 • SELECT * FROM filmboxoffice.film_news;

Result Grid Filter Rows: Edit: Export/Import: Wrap Cell Content:

	NEWS_ID	NEWS_TITLE	NEWS_CONTENT	ORIGINAL_URL	PUBLISH
▶	1	中国泛娱乐创新峰会“升级·未来电影”会场即...	中国泛娱乐创新峰会“升级·未来电...	http://www.entgroup.cn/news/Exclusive/14373...	2016-11-14
	2	双11狂欢夜背后 一边控制钱包一边控制快乐	双11狂欢夜背后 一边控制钱包一...	http://www.entgroup.cn/news/Exclusive/11373...	2016-11-14
	3	重磅代表为项目站台 寻找未来影视新锐	重磅代表为项目站台 寻找未来影...	http://www.entgroup.cn/news/Exclusive/18374...	2016-11-14
	4	ENAAwards最佳娱乐营销案例优秀入围推荐 (...	ENAAwards最佳娱乐营销案例优秀...	http://www.entgroup.cn/news/Exclusive/18374...	2016-11-14
	5	精品网络剧强力反哺一线卫视 台网壁垒再...	精品网络剧强力反哺一线卫视 台...	http://www.entgroup.cn/news/Exclusive/14373...	2016-11-14
	6	三大视频网站付费会员实现4-5倍增长 自制...	三大视频网站付费会员实现4-5倍...	http://www.entgroup.cn/news/Exclusive/11373...	2016-11-14
	7	从《西部世界》回看今日，大数据会是下一...	从《西部世界》回看今日，大数...	http://www.entgroup.cn/news/Exclusive/08373...	2016-11-04
	8	“IP买卖”热潮中 怎样才能让你的IP卖更多钱	“IP买卖”热潮中 怎样才能让你的IP...	http://www.entgroup.cn/news/Exclusive/09373...	2016-11-04
	9	2016中国泛娱乐指数盛典专家评审会召开	2016中国泛娱乐指数盛典专家评...	http://www.entgroup.cn/news/Exclusive/10373...	2016-11-14
	10	400亿指标完成国产占比近六成 全年增幅或...	400亿指标完成国产占比近六成 ...	http://www.entgroup.cn/news/Exclusive/10373...	2016-11-14
	11	极客当道：化身“天使”守护原创者“顶层设...	极客当道：化身“天使”守护原创者...	http://www.entgroup.cn/news/Exclusive/15374...	2016-11-14
	12	有钱，有项目，先导概念片你来拍！	有钱，有项目，先导概念片你来...	http://www.entgroup.cn/news/Exclusive/11373...	2016-11-14
	13	中南文化以收购业绩成效明显 文化产业...	中南文化以收购业绩成效明显 ...	http://www.entgroup.cn/news/Exclusive/27371...	2016-10-24
	14	双11狂欢夜背后 一边控制钱包一边控制快乐	双11狂欢夜背后 一边控制钱包一...	http://www.entgroup.cn/news/Exclusive/11373...	2016-11-14
	15	多家重量级片方集聚 时空引力场即将引爆	多家重量级片方集聚 时空引力场...	http://www.entgroup.cn/news/Exclusive/02372...	2016-11-04
	16	还在看不起网大吗？可能它们就是未来的电影	还在看不起网大吗？可能它们就...	http://www.entgroup.cn/news/Exclusive/07372...	2016-11-04
	17	《蒙面唱将猜猜猜》全面升级 引爆Q4收视...	《蒙面唱将猜猜猜》全面升级 引...	http://www.entgroup.cn/news/Exclusive/26371...	2016-10-24
	18	ENAAwards最佳娱乐营销案例优秀入围推荐 (...	ENAAwards最佳娱乐营销案例优秀...	http://www.entgroup.cn/news/Exclusive/18374...	2016-11-14

Web Crawler

Results

名称	修改日期	类型	大小
.news.Markets.3055425.shtml.txt	2018/12/13 11:31	文本文档	36 KB
.news.Markets.3056711.shtml.txt	2018/12/13 11:30	文本文档	36 KB
.news.Markets.3056715.shtml.txt	2018/12/13 11:30	文本文档	36 KB
.news.Markets.3056737.shtml.txt	2018/12/13 11:30	文本文档	36 KB
.news.Markets.3057950.shtml.txt	2018/12/13 11:28	文本文档	36 KB
.news.Markets.3057951.shtml.txt	2018/12/13 11:28	文本文档	36 KB
.news.Markets.3057953.shtml.txt	2018/12/13 11:28	文本文档	36 KB
.news.Markets.3057956.shtml.txt	2018/12/13 11:28	文本文档	36 KB
.news.Markets.3057957.shtml.txt	2018/12/13 11:28	文本文档	36 KB
.news.Markets.3154062.shtml.txt	2018/12/13 11:33	文本文档	35 KB
.news.Markets.3154063.shtml.txt	2018/12/13 11:33	文本文档	35 KB
.news.Markets.3154068.shtml.txt	2018/12/13 11:33	文本文档	35 KB
.news.Markets.3154069.shtml.txt	2018/12/13 11:33	文本文档	35 KB
.news.Markets.3154070.shtml.txt	2018/12/13 11:33	文本文档	35 KB
.news.Markets.3156757.shtml.txt	2018/12/13 11:30	文本文档	36 KB
.news.Markets.3156759.shtml.txt	2018/12/13 11:30	文本文档	36 KB
.news.Markets.3156763.shtml.txt	2018/12/13 11:30	文本文档	36 KB
.news.Markets.3156765.shtml.txt	2018/12/13 11:30	文本文档	36 KB
.news.Markets.3156766.shtml.txt	2018/12/13 11:30	文本文档	36 KB
76.0.txt	2018/12/13 11:27	文本文档	35 KB
76.1.txt	2018/12/13 11:28	文本文档	36 KB
76.2.txt	2018/12/13 11:28	文本文档	38 KB
76.3.txt	2018/12/13 11:28	文本文档	37 KB
76.4.txt	2018/12/13 11:28	文本文档	36 KB
76.5.txt	2018/12/13 11:29	文本文档	36 KB
76.6.txt	2018/12/13 11:29	文本文档	36 KB
76.7.txt	2018/12/13 11:29	文本文档	35 KB
76.8.txt	2018/12/13 11:29	文本文档	37 KB
76.9.txt	2018/12/13 11:30	文本文档	36 KB
76.10.txt	2018/12/13 11:30	文本文档	36 KB
76.11.txt	2018/12/13 11:30	文本文档	36 KB
76.12.txt	2018/12/13 11:30	文本文档	36 KB

```
Run: 07EntgroupExclusiveMainWebCrawler ×
"C:\Program Files\Python36\python.exe" "D:/SISU/Courses/2018计算新闻学(研)/Lecture 6/07EntgroupExclusiveMainWebCrawler.py"
百度音乐: 起个大早赶个晚集 : /news/Markets/1358623.shtml 2018-12-13 11:27:41
1 file saved!
Next...
跟《国家宝藏》学做历史剧, 安全上星又有市场! : /news/Markets/1358622.shtml 2018-12-13 11:27:42
1 file saved!
Next...
2019或成中国科幻电影元年, 作品质量有待观察 : /news/Markets/1358618.shtml 2018-12-13 11:27:43
1 file saved!
Next...
年末将至, 来看看近期的13部献礼剧 : /news/Markets/1358617.shtml 2018-12-13 11:27:43
1 file saved!
Next...
近三分之一的直播平台倒闭, 不少中小平台开始选择出海 : /news/Markets/1258555.shtml 2018-12-13 11:27:44
1 file saved!
Next...
文娱快讯|漫威将拍《奇异博士2》导演回归; 赵薇&田壮壮合作首部话剧《求证》 : /news/Markets/1258553.shtml 2018-12-13 11:27:45
1 file saved!
Next...
总结2018时, 我们看到了些许中国动画电影的希望 : /news/Markets/1258551.shtml 2018-12-13 11:27:45
1 file saved!
Next...
星美回应倒闭传闻 公司近期将回归正轨 : /news/Markets/1258537.shtml 2018-12-13 11:27:46
1 file saved!
```





The End of Lecture 9

Thank You

<http://www.wangting.ac.cn>

