



New Media Data Analytics and Application

Lecture 7: Information Acquisition
An Integration

Ting Wang

- Product-Oriented Data Collection
- Make a Web Crawler
- System Integration





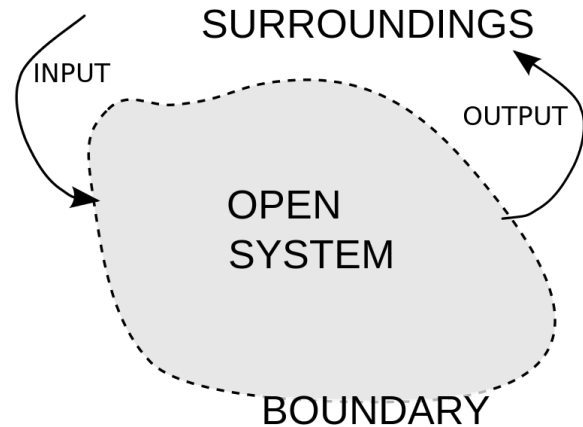
You should know your purpose of data collection

Product-Oriented Data Collection

Product-Oriented Data Collection

Firstly, Product is the most important.

- It should be a system.



Product-Oriented Data Collection

Secondly, all the parts of the system should be designed as a product.

- Data from API are no longer a part of the original system, they belong to your new system



Product-Oriented Data Collection

Thirdly, different data should be fused and stored together, just as a database for a product

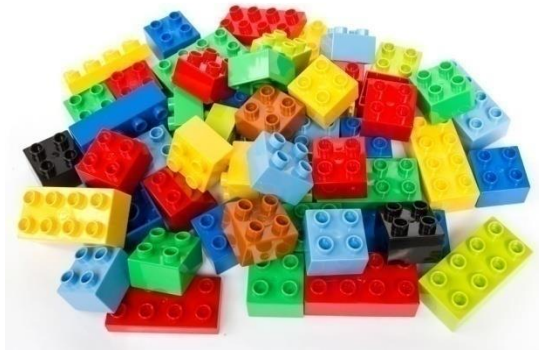
- After data fusion, all the data look as if they are collected from one source, not different.



Product-Oriented Data Collection

Fourthly, modular construction is important.

- Understanding the function of every modular, and integrating them!



Product-Oriented Data Collection

Last but not the least, do NOT want to design a very huge system in the very beginning, that is impossible.

- Rome was not built in a day.





to make a web spider for your projects

Web Crawler

Objectives: Data Collection

If there is no API provided by the website, you may use Web Crawler.

You should ask yourself:

- What is your purpose?
- Which website is the easiest for web scraping?



Analyze the Structure of the Potential Candidate Website

- Employ a good web explorer

Eg. Google Chrome



Google chrome

- Select a webpage as a start

Eg. <http://www.entgroup.cn/news/Exclusive/> or

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

Web Crawler

<http://www.entgroup.cn/news/Exclusive/>

The screenshot shows the website www.entgroup.cn/news/Exclusive/. The header includes navigation links for ENGLISH, 登录 (Login), 注册 (Register), and 公司介绍 (Company Introduction). Below the header is a purple banner for 艺恩网 (Entgroup.cn) with a search bar. The main content area features a large article titled 《相爱相亲》三性统一，获不同层次观众满意评价 (Love and Family: Three in One, Satisfying Different Levels of Audience). The article includes a photo of the main cast and text describing the film's success. To the right of the article is a sidebar with '热门文章' (Popular Articles) and '热门标签' (Hot Tags). The footer contains a section for '2017 中国演艺赛事数据报告' (2017 China Performance and Entertainment Industry Data Report).

艺恩网
entgroup.cn

分类 艺恩智库 中国票房 交易平台

请输入您感兴趣的关键词

艺恩数据
VERSION 2.1

工作台全新上线

立即下载

原创 政策 人物 跨界 资本 营销 市场

热门文章 本周 本月

《相爱相亲》三性统一，获不同层次观众满意评价

张艺谋导演的新作《相爱相亲》是中国电影观众满意度调查在2017年10月底到11月份调查的唯一一部爱情电影。

2017-12-11 电影 观众满意度

艺恩IP中心推荐：《情梅》三部曲，诡异奇案窥探人性深处的道德底线

都市悬疑、商战类作品《情梅》三部曲，为我们讲述一个诡异的奇案，揭露出一段不为人知的爱恨纠葛。

2017-12-11 版权 IP

12月《中国演艺赛事数据报告》

以西十区独有的交易数据为基础，联手艺恩打造的《中国演艺赛事数据报告》又

艺恩IP中心推荐：《情梅》三部曲，诡异奇案窥探人性深处的道德底线

艺恩IP中心推荐：《毒妃倾城：嚣张王爷浪荡点》能否成为下一部《锦绣未央》

艺恩IP中心推荐：《我们都不擅长告别》，而我却要送你离开

艺恩IP中心推荐：《猫鼠计划》，上演高智商版《伪装者》

艺恩IP中心推荐：《试用期千金》，上演甜蜜、发糖版《妻子的谎言》

热门标签 本周 本月

IP 改编 泛娱乐

制片人 银幕 影视

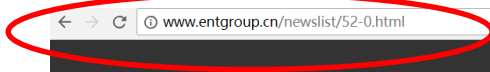
[Notes:]

For Class 2014, I employed this web page.
However, it seems that this page has not been updated for a long time.



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

- 1
- 2
- 3
- 4
- 5
- 下一页>



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 on the right. 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 rules:

```
body {  
  background-color: transparent;  
  margin: 0px;  
  height: 100%;  
}
```

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



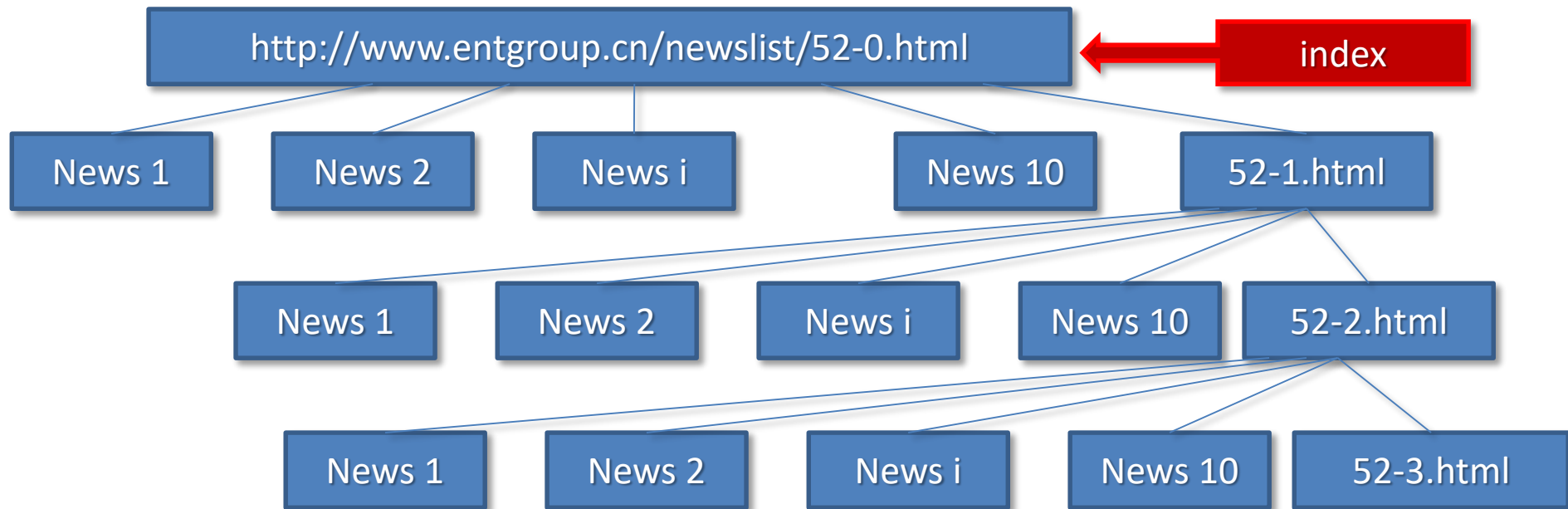
Ask A Question

What is your Web Scraping Strategy of News List?



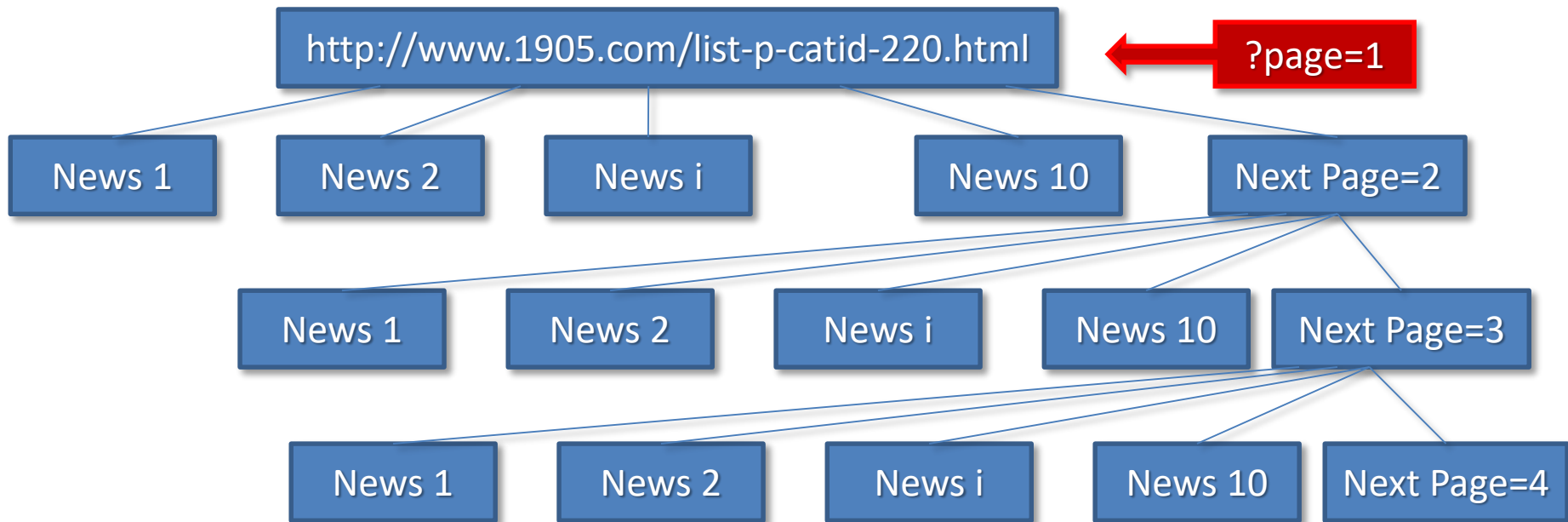
Web Crawler

Eg.1 Solutions to Web Scraping Strategy of EntGroup News



Web Crawler

Eg.2 Solutions to Web Scraping Strategy of M1905 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)
```



Evolution based on the Prototype

```
import urllib.request
response = urllib.request.urlopen('http://www.entgroup.cn/newslist/52-0.html')
HTMLText = response.read()

with open('WebCrawler/entgroup/Exclusive/52.0.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?



Beautifulsoup

Beautiful Soup is a Python library designed for quick turnaround projects like screen-scraping(<https://www.crummy.com/software/BeautifulSoup/bs4/doc.zh/index.html>)

Beautiful Soup 4.4.0 documentation » index

Table Of Contents

- Beautiful Soup
- Documentation
 - Getting help
- Quick Start
- Installing Beautiful Soup
 - Problems after installation
 - Installing a parser
- Making the soup
 - Making the soup
 - Making the soup
- Tag
 - Name
 - Attributes
 - Multi-valued attributes
- NavigableString
- BeautifulSoup
- Comments and other special strings
- Navigating the tree
 - Going down
 - Navigating using tag names
 - .contents and .children
 - .descendants
 - .string
 - .strings and .stripped_strings

Beautiful Soup Documentation

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

These instructions illustrate all major features of Beautiful Soup 4, with examples. I show you what the library is good for, how it works, how to use it, how to make it do what you want, and what to do when it violates your expectations.

The examples in this documentation should work the same way in Python 2.7 and Python 3.2.

You might be looking for the documentation for [Beautiful Soup 3](#). If so, you should know that Beautiful Soup 3 is no longer being developed, and that Beautiful Soup 4 is recommended for all new projects. If you want to learn about the differences between Beautiful Soup 3 and Beautiful Soup 4, see [Porting code to BS4](#).

This documentation has been translated into other languages by Beautiful Soup users:

- 这篇文档当然还有中文版.
- このページは日本語で利用できません(外部リンク)
- 이 문서는 한국어 번역도 가능합니다. (외부 링크)



Getting help

If you have questions about Beautiful Soup, or run into problems, [send mail to the discussion group](#). If your problem involves parsing an HTML document, be sure to mention what the `diagnose()` function says about that document.

Quick Start

Here' s an HTML document I' ll be using as an example throughout this document. It' s part of a story from *Alice in Wonderland*.

Web Scraping using BeautifulSoup

- Installation:

```
pip install beautifulsoup4
```

```
C:\Windows\system32>pip3 install beautifulsoup4
Collecting beautifulsoup4
  Downloading https://files.pythonhosted.org/packages/9e/d4/10f46e5cfac773e22707237bfcd51bbffeaf0a576b0a847ec7ab15bd7ace/beautifulsoup4-4.6.0-py3-none-any.whl (86kB)
    100% |#####| 92kB 209kB/s
Installing collected packages: beautifulsoup4
Successfully installed beautifulsoup4-4.6.0
```



find()

- Look for the first one

findAll()

- Look for all

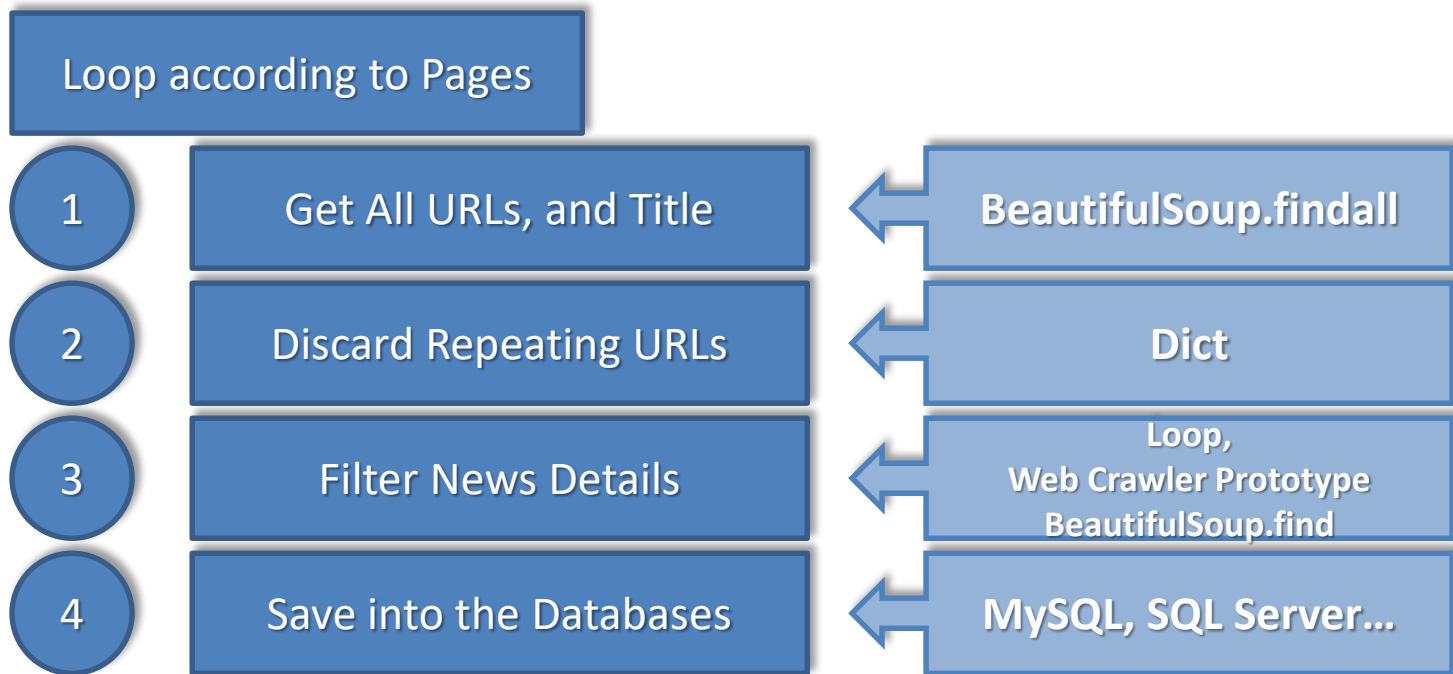
Ref.

<http://www.jb51.net/article/65287.htm>



Web Crawler

- Steps:



- 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/52-0.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/Exclusive/', 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/Exclusive/52.0.txt', 'wb') as f:
    f.write(HTMLText)
```



• Results of Step 1.

Run: 03EntgroupExclusiveMainWebCrawStep1 (1) x

"C:\Program Files\Python36\python.exe" "D:/SISU/Courses/2018新媒体数据:/news/Exclusive/2345585.shtml

/news/Exclusive/2345585.shtml

艺恩IP中心推荐: 小说《青春是一个人的兵荒马乱》

/news/Exclusive/2345584.shtml

/news/Exclusive/2345584.shtml

文娱12点: 腾讯影业成立腾影发行公司; 电影《昼颜》定档5月18日

/news/Exclusive/2345580.shtml

/news/Exclusive/2345580.shtml

优酷春集发百部新作, 你有多少睡眠时间给了圈层内容?

/news/Exclusive/2245578.shtml

/news/Exclusive/2245578.shtml

文娱快讯|冯小刚《芳华》拿下导演协会年度影片; 黄子韬为腾讯微视首位代言人

← → ↻ www.entgroup.cn/newslist/52-0.html

原创 政策 人物 跨界 资本 营销 市场

 艺恩IP中心推荐: 小说《青春是一个人的兵荒马乱》
小说时代背景为现代都市, 快消品、时装品牌、数码电子产品等可大量植入。
艺恩IP中心 2018-04-23 改编 IP 0

 文娱12点: 腾讯影业成立腾影发行公司; 电影《昼颜》定档5月18日
艺恩为您带来文娱产业每日最新资讯。
mary 2018-04-23 泛娱乐 0

 优酷春集发百部新作, 你有多少睡眠时间给了圈层内容?
优酷春集在北京奥林匹克中心举办, 平台战略发布的同时公布了上百档即将上线的大剧、热综。
李敬蕊 2018-04-23 优酷 0

 文娱快讯|冯小刚《芳华》拿下导演协会年度影片; 黄子韬为腾讯微视首位代言人
艺恩为您带来文娱产业最新资讯。



- 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/52-0.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/Exclusive/', 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/Exclusive/52.0.txt', 'wb') as f:
    f.write(HTMLText)
print(URLdict)
```





- 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/52-0.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/Exclusive/', 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/Exclusive/52.0.txt', 'wb') as f:
    f.write(HTMLText)
```



- Results of Step 3.



```
Run: 05EntgroupExclusiveMainWebCrawStep3 x
"C:\Program Files\Python36\python.exe" "D:/SISU/Courses/2018新媒体数据分析与应用（2015网新）/Lecture 7/05EntgroupExclusiveMainWebCrawStep3.py"
艺恩IP中心推荐：小说《青春是一个人的兵荒马乱》：/news/Exclusive/2345585.shtml

艺恩IP中心推荐：小说《青春是一个人的兵荒马乱》

标签：
    IP改编
来源：艺恩网作者：艺恩IP中心2018-04-23

本文为艺恩网原创专稿，尊重作者版权，转载请注明出处及作者
[摘要]

小说时代背景为现代都市，快消品、时装品牌、数码电子产品等可大量植入。

我们会错过一趟列车，一场烟火，一次机会，甚至是错过你。可那些年隐匿于岁月里的青春故事，没有如果，只有结果……艺恩IP中心今日推荐，青春、爱情类型小说《青春是一个人的兵荒马乱》，讲述陈默，一个蹩脚的三流小说码字员，通过他的笔触，讲述了大学同学，

编辑：mary

文娱12点：腾讯影业成立腾影发行公司；电影《昼颜》定档5月18日：/news/Exclusive/2345584.shtml

文娱12点：腾讯影业成立腾影发行公司；电影《昼颜》定档5月18日
```

- 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/52-'+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/Exclusive/', 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/Exclusive/52."+str(i)+'.txt', 'wb') as f:
        f.write(HTMLText)
```



• Results of Step 4.

```
Run: 06EntgroupExclusiveMainWebCrawStep4 x
编辑: mary

亿元投入成标配，网综的“大片时代”来了？ : /news/Exclusive/0344866.shtml

亿元投入成标配，网综的“大片时代”来了？

标签:
    网络综艺播放量
来源: 艺恩网作者: 耿耀2018-02-03

本文为艺恩网原创专稿，尊重作者版权，转载请注明出处及作者
[摘要]

网综在去年开始二次升级后，2018年将全面进入大片时代。“大片时代背后是网综在投入、制作、商业上的全面升级。”

根据艺恩发布的《2017中国网络综艺市场白皮书》显示，2017年，主流视频网站共上线159部网络综艺，投资规模达43亿，同比增长43%，播放量超500亿，市场表现与电视综艺比肩。在业内人士看来，随着蓬勃发展，网综与台综的界限正进一步被打破。日前在上海举行的优

编辑: yvonne

Process finished with exit code 0
```

名称	修改日期	类型	大小
52.0.txt	2018/4/23 22:36	文本文档	35 KB
52.1.txt	2018/4/23 22:36	文本文档	36 KB
52.2.txt	2018/4/23 22:37	文本文档	35 KB
52.3.txt	2018/4/23 22:37	文本文档	35 KB
52.4.txt	2018/4/23 22:37	文本文档	35 KB
52.5.txt	2018/4/23 22:38	文本文档	35 KB
52.6.txt	2018/4/23 22:38	文本文档	36 KB
52.7.txt	2018/4/23 22:38	文本文档	36 KB
52.8.txt	2018/4/23 22:39	文本文档	36 KB
52.9.txt	2018/4/23 22:39	文本文档	36 KB



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)						
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),
    PUBLISH_DATE     DATE
);
```



Store into Databases

1. Discard Repetitions

```
sqlstr="select * from film_news where ORIGINAL_URL='http://www.entgroup.cn'+link+'"'
cursor.execute(sqlstr)
numrows = len(cursor.fetchall())
if numrows>0:
    continue; #判断相同URL是否已经在数据库中存在, 若存在, 则下一轮
else:
    # 判断相同URL是否已经在数据库中存在, 若不存在, 则取数据, 并添加至数据库
```



Store into Databases

2. Discard Repetitions

```
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.Exclusive.1745525.shtml.txt	2018/4/23 23:31	文本文档	36 KB
.news.Exclusive.1745526.shtml.txt	2018/4/23 23:31	文本文档	36 KB
.news.Exclusive.1745528.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.1845531.shtml.txt	2018/4/23 23:31	文本文档	36 KB
.news.Exclusive.1845539.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.1845540.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.1945541.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.1945549.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.1945551.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.1945554.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.1945555.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.1945557.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2045562.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2045564.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2045566.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2145569.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2145570.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2145571.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2245574.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2245577.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2245578.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2345580.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2345584.shtml.txt	2018/4/23 23:31	文本文档	35 KB
.news.Exclusive.2345585.shtml.txt	2018/4/23 23:31	文本文档	35 KB
52.0.txt	2018/4/23 23:31	文本文档	35 KB
52.1.txt	2018/4/23 23:31	文本文档	36 KB
52.2.txt	2018/4/23 23:31	文本文档	35 KB

```
Run: 07EntgroupExclusiveMainWebCrawler.py
"C:\Program Files\Python36\python.exe" "D:\SISU\Courses\2018新媒体数据分析与应用(2015网新)/Lecture 7/07EntgroupExclusiveMainWebCrawler.py"
艺恩IP中心推荐: 小说《青春是一个人的兵荒马乱》 : /news/Exclusive/2345585.shtml 2018-04-23 23:31:18
文娱12点: 腾讯影业成立腾影发行公司; 电影《巨额》定档5月18日 : /news/Exclusive/2345584.shtml 2018-04-23 23:31:18
优酷春集发百部新作, 你有多少睡眠时间给了圈层内容? : /news/Exclusive/2345580.shtml 2018-04-23 23:31:18
文娱快讯|冯小刚《芳华》拿下导演协会年度影片; 黄子韬为腾讯微视首位代言人 : /news/Exclusive/2245578.shtml 2018-04-23 23:31:18
艺恩IP中心推荐: 小说《无间英雄》, 警察的传奇人生 : /news/Exclusive/2245577.shtml 2018-04-23 23:31:19
《最后的锦衣卫》导演杨振豪: 在国际舞台上讲好中国故事 : /news/Exclusive/2245574.shtml 2018-04-23 23:31:19
文娱快讯|《画皮3》宣布启动电影/网剧; 马化腾入选《财富》2018全球最伟大50位领袖 : /news/Exclusive/2145571.shtml 2018-04-23 23:31:19
艺恩IP中心推荐: 小说《孤星满月》, 人生若只如初见, 何事悲风秋画扇 : /news/Exclusive/2145570.shtml 2018-04-23 23:31:19
艺恩对话|Base FX副总裁谢宁: 影视后期需要“老鹰”不需要肯德基的“肉鸡” : /news/Exclusive/2145569.shtml 2018-04-23 23:31:20
北影节|孙向辉: 观众满意度调查引导市场新动能 : /news/Exclusive/2045566.shtml 2018-04-23 23:31:20
承华传媒“拿下”《极限特工4》 合资公司内地牵手花儿影视 : /news/Exclusive/2045564.shtml 2018-04-23 23:31:20
文娱12点:《头号玩家》延至5月29日上映;《妖猫传》被指抄袭 : /news/Exclusive/2045562.shtml 2018-04-23 23:31:20
淘票票公布2018战略规划 发布电影宣发平台“灯塔” : /news/Exclusive/1945557.shtml 2018-04-23 23:31:21
张昭:“屌丝”转中产将成为中国电影产业升级的价值蓝海 : /news/Exclusive/1945555.shtml 2018-04-23 23:31:21
艺恩把脉中国电影市场趋势 分享艺恩数据APP最新亮点 : /news/Exclusive/1945554.shtml 2018-04-23 23:31:21
北影节|艺恩论坛: 版权方、制作方、资方畅谈电影未来趋势 : /news/Exclusive/1945551.shtml 2018-04-23 23:31:21
文娱12点: 黄渤导演处女作定名《一出好戏》定档8.10; 中央广播电视总台正式揭牌亮相 : /news/Exclusive/1945549.shtml 2018-04-23 23:31:22
艺恩IP中心推荐: 搞笑类型漫画《我的不靠谱王子殿下》 : /news/Exclusive/1845540.shtml 2018-04-23 23:31:22
2022年票房将超900亿 如何把握行业未来增长点 : /news/Exclusive/1845539.shtml 2018-04-23 23:31:22
文娱12点:《黄金花》改档4.28; 腾讯首次布局游戏MCN机构 : /news/Exclusive/1845539.shtml 2018-04-23 23:31:22
老神纪录被新神“秒破”, 谁是一个超级IP? : /news/Exclusive/1045468.shtml 2018-04-23 23:31:23
首都之星艺术影院联盟建立 首批20家影院每周不少于两场艺术影片放映 : /news/Exclusive/1745528.shtml 2018-04-23 23:31:23
文娱12点|斯皮尔伯格个人票房突破100亿美元; 赵薇夫妇入证证券市场5年 : /news/Exclusive/1745524.shtml 2018-04-23 23:31:23
0. 0%
艺恩IP中心推荐:《糖糖》玫瑰与面包, 你会选择哪一个? : /news/Exclusive/1845531.shtml 2018-04-23 23:31:24
```



Tips:

- Web Crawler is very complex. If you want to use web Crawler, you should build them individually for each different websites.





build an information collection system by APIs and web crawlers

System Integration

System Integration

API Prototype

Do you remember
Weather Forecast?

1. Front Page

2. Condition Input

3. API

```
4 from flask import Flask
5 from flask import request
6 import urllib
7
8 app = Flask(__name__)
9
10 @app.route('/', methods=['GET', 'POST'])
11 def home():
12     return '''<h1>Home</h1>
13         <p><a href="/CityWeather">Weather Forecast</a></p>'''
14
15 @app.route('/CityWeather', methods=['GET'])
16 def login_form():
17     return '''<form action="/CityWeather" method="POST">
18         <p>City Name <input name="CityName"></p>
19         <p><button type="submit">Weather</button></p>
20         </form>'''
21
22 @app.route("/CityWeather", methods=['POST'])
23 def login():
24     # get data from Baidu:
25     url = 'http://apis.baidu.com/heweather/weather/free?city=' + request.form['CityName']
26     req = urllib.request.Request(url)
27     req.add_header("apikey", "5e7ef01f44164283876b0dd0cbd1461e") # Dont forget to change your spikye
28     resp = urllib.request.urlopen(req)
29     content = resp.read()
30     if (content):
31         return content
32     return '<h3>Bad City Name.</h3>'
33
34 if __name__ == '__main__':
35     app.run()
```



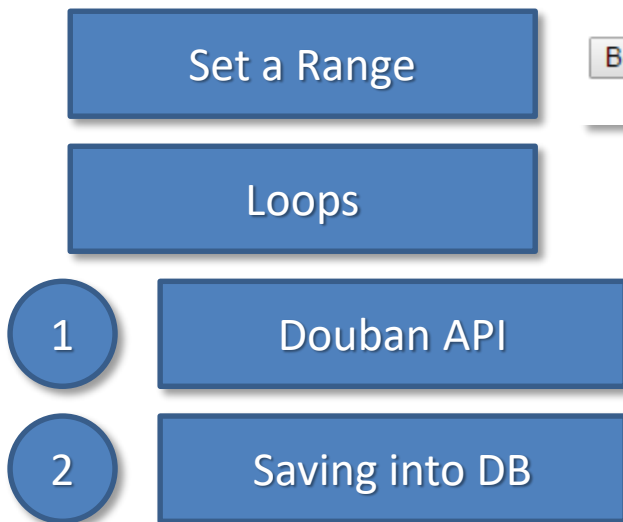
API for Douban

```
# -*- coding: utf-8 -*-  
import sys,urllib.request,json  
  
url = 'https://api.douban.com/v2/movie/subject/1764730' #豆瓣电影，修改或循环遍历那个数字  
  
req = urllib.request.Request(url)  
  
req.add_header("apikey", "5e7ef01f44164283876b0dd0cbd1461e")#换上你自己的API Key  
  
resp = urllib.request.urlopen(req)  
response = resp.read().decode('utf-8')  
content = json.dumps(json.loads(response),ensure_ascii=False)  
if(content):  
    print(content)
```



System Integration

Steps and Results



← → ↻ ⓘ 127.0.0.1:5000/FilmBoxOffice

Douban Film ID

From: To:

← → ↻ ⓘ 127.0.0.1:5000/FilmBoxOffice

Searching Finished, totally 5 films

```
127.0.0.1 - - [22/Nov/2016 23:52:54] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [22/Nov/2016 23:52:55] "GET /FilmBoxOffice HTTP/1.1" 200 -
"Est": "2012"
"奇异博士": "2016"
"情迷黑森林": "2005"
"追忆往事": "1987"
"最后的铃声": "1989"
127.0.0.1 - - [22/Nov/2016 23:53:38] "POST /FilmBoxOffice HTTP/1.1" 200 -
127.0.0.1 - - [22/Nov/2016 23:55:29] "GET /FilmBoxOffice HTTP/1.1" 200 -
```



System Integration

Results in Database

2 • `select * from film_info`

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

NS_COUNT	SHARE_URL	SUBTYPE	SUMMARY	TITLE	WISH_COUNT	RELEASE_YEAR
	"https://m.douban.com/movie/subject/3025363"	"movie"	"	"Est"	0	2012
	"https://m.douban.com/movie/subject/3025375"	"movie"	"出色的神经外科医生斯蒂芬·斯特兰奇（本...	"奇异博士"	24972	2016
	"https://m.douban.com/movie/subject/3025376"	"tv"	"最近，在甜品界中出现了一个“罗宾汉”，常...	"情迷黑森林"	150	2005
	"https://m.douban.com/movie/subject/3025379"	"movie"	"片名：似曾相识/往事重现 Vec viđeno (Dějà...	"追忆往事"	70	1987
	"https://m.douban.com/movie/subject/3025393"	"movie"	"	"最后的铃声"	3	1989
*	NULL	NULL	NULL	NULL	NULL	NULL



System Integration



Ask A Question

How to integrate Film Box Office with the data from web APIs and crawlers?



Film Box Office Prediction

- Add Historical Film Box Office to Databases
- Data Clean and Natural Language Processing
- Use Statistical or Machine Learning methods

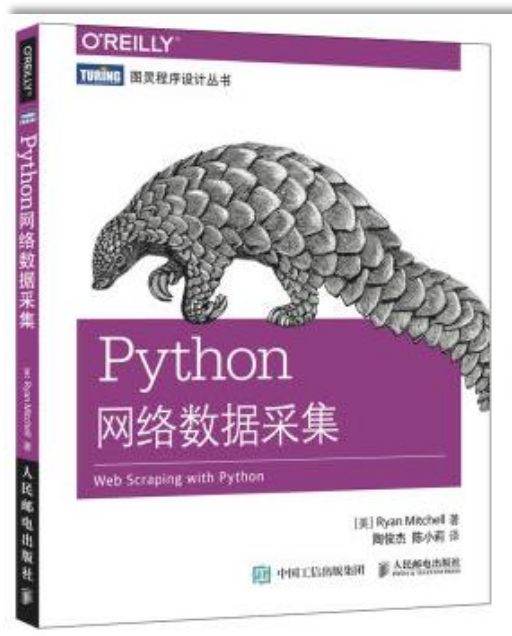




上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Reference

Python网络数据采集



Python中使用Beautiful Soup库的超详细教程

- <http://www.jb51.net/article/65287.htm>





上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Homework

Homework

1. Finish the Web Crawler, and save the data into the Data Base.
(It will be check in **May 8.**)
2. Try to build a web crawler for your group





The End of Lecture 7

Thank You

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

