上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

# New Media
# Data Analytics and Application

## Lecture 7: Information Acquisition
## An Integration

Ting Wang

# Outlines

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


上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

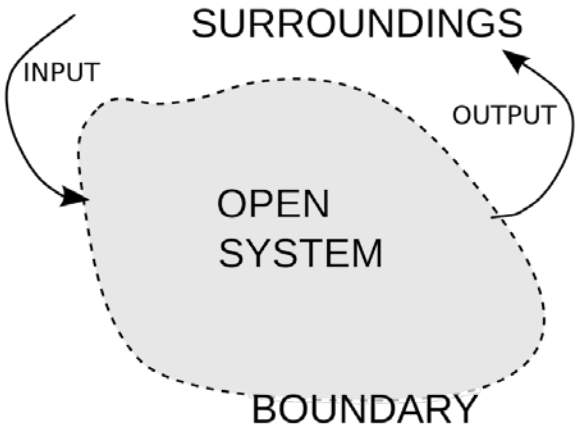
You should know your purpose of data collection

# 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

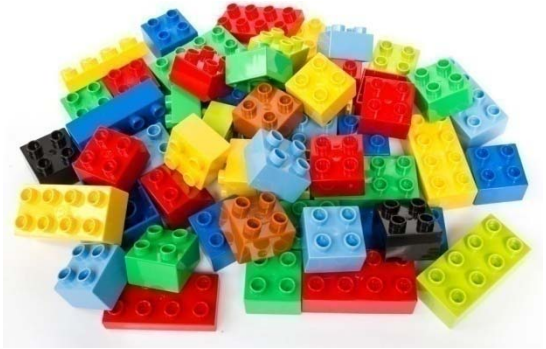
***Thirdly, different data should be fused and stored together, just as a database for a product***

- 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.


上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

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?

# Web Crawler

## *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/

# Web Crawler

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

# Web Crawler


*Why we can choose* http://www.entgroup.cn/news/Exclusive/

*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>

LIST

Next Pages

www.entgroup.cn/news/Exclusive/

### 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，是什么让资方肯下重金布局的？

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，上星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

虽然有时会嘟着“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

### 有钱，有项目，先导概念片你来拍！

上海外国语大学 SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

# Web Crawler

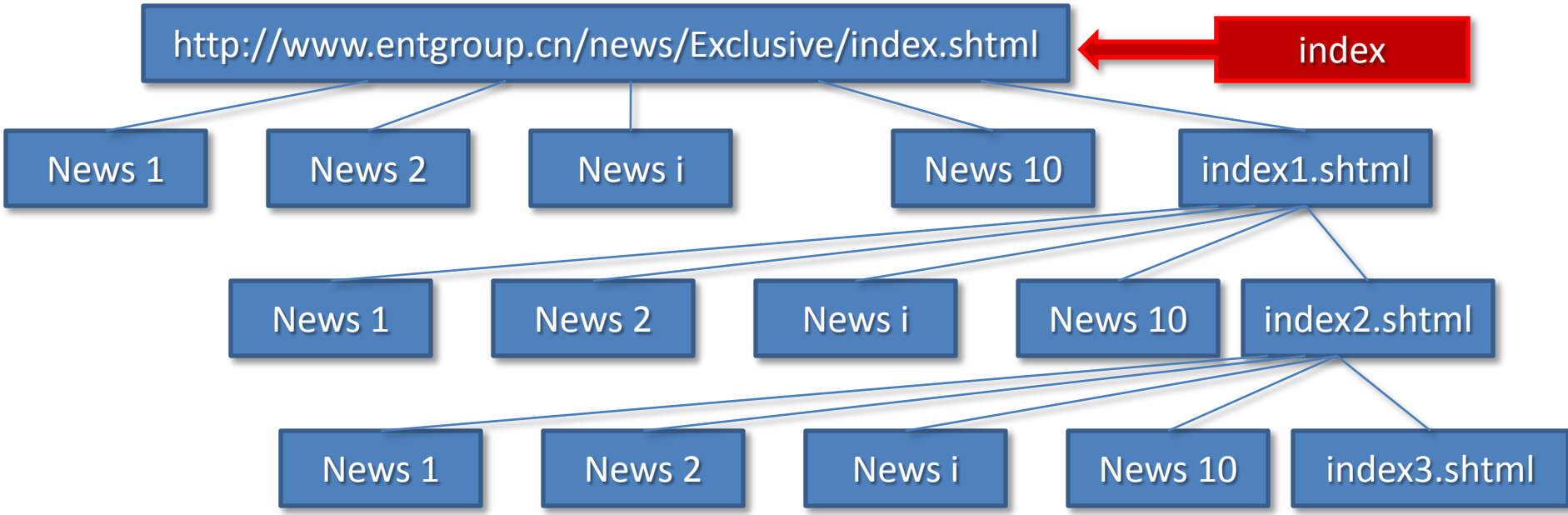
## 更多工具 开发者工具

# Web Crawler


***What is your Web Scraping Strategy of News List?***

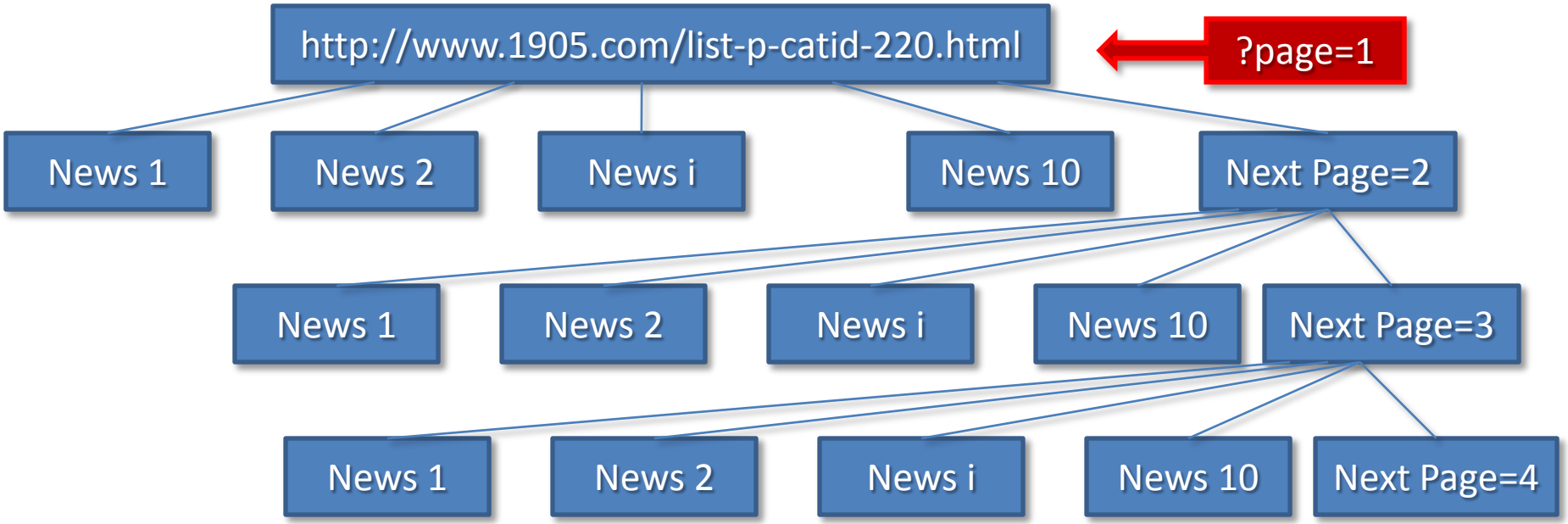
# Web Crawler

## *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/news/Exclusive/index.shtml')
HTMLText = response.read()

with open('WebCrawler/entgroup/Exclusive/index.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/

You didn't write that awful page. You're just trying to get some data out of it. Beautiful Soup is here to help. Since 2004, it's been saving programmers hours or days of work on quick-turnaround screen scraping projects.

## Beautiful Soup

"A tremendous boon." -- Python411 Podcast

[ Download | Documentation | Hall of Fame | Source | Discussion group ]

If Beautiful Soup has saved you a lot of time and money, the best way to pay me back is to check out *Constellation Games*, my sci-fi novel about alien video games.
You can read the first two chapters for free, and the full novel starts at 5 USD. Thanks!

*If you have questions, send them to the discussion group. If you find a bug, file it.*

Beautiful Soup is a Python library designed for quick turnaround projects like screen-scraping. Three features make it powerful:

1. Beautiful Soup provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree: a toolkit for dissecting a document and extracting what you need. It doesn't take much code to write an application
2. Beautiful Soup automatically converts incoming documents to Unicode and outgoing documents to UTF-8. You don't have to think about encodings, unless the document doesn't specify an encoding and Beautiful Soup can't detect one. Then you just have to specify the original encoding.
3. Beautiful Soup sits on top of popular Python parsers like lxml and html5lib, allowing you to try out different parsing strategies or trade speed for flexibility.

Beautiful Soup parses anything you give it, and does the tree traversal stuff for you. You can tell it "Find all the links", or "Find all the links of class externalLink", or "Find all the links whose urls match "foo.com", or "Find the table heading that's got bold text, then give me that text."

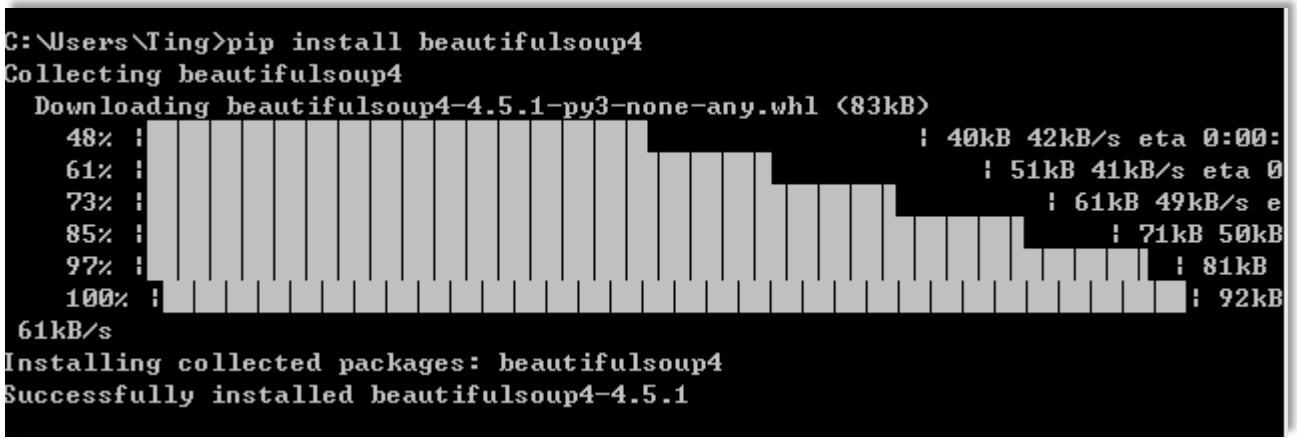
Valuable data that was once locked up in poorly-designed websites is now within your reach. Projects that would have taken hours take only minutes with Beautiful Soup.

Interested? Read more.

# *Web Scraping using Beautifulsoup*

- Installation:

Please make sure that you have installed ez_setup.py, if you not sure , please review Lecture 4



# Web Crawler

## *find()*

- Look for the first one

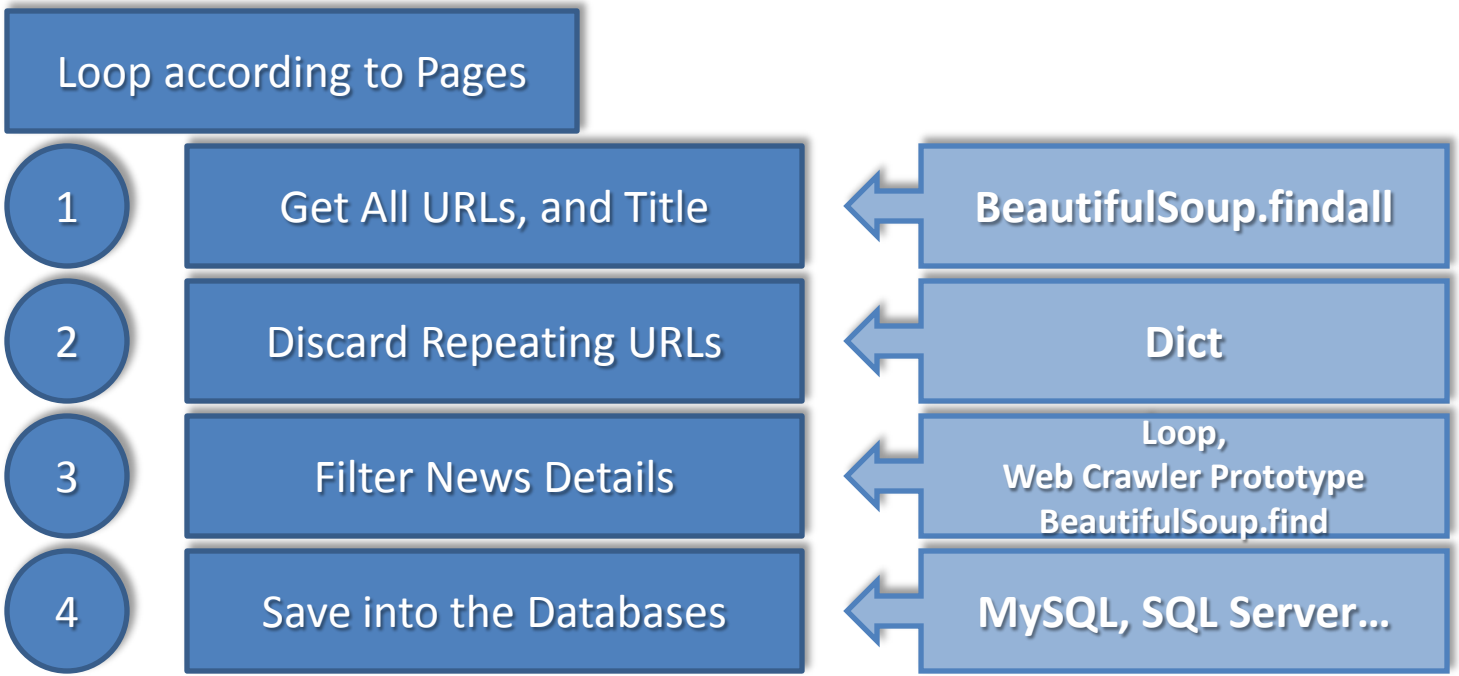
## *findAll()*

- Look for all

***Ref.***
***http://www.jb51.net/article/65287.htm***

# Web Crawler

- Steps:



# Web Crawler

- Step 1 : Get All URLs, and Title

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

response = urllib.request.urlopen('http://www.entgroup.cn/news/Exclusive/index.shtml')
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/index.txt', 'wb') as f:
    f.write(HTMLText)
```

- 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/news/Exclusive/index.shtml')
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/index.txt', 'wb') as f:
   f.write(HTMLText)
print(URLdict)
```

# Web Crawler

- 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/news/Exclusive/index.shtml')
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/index.txt', 'wb') as f:
    f.write(HTMLText)
```

# Web Crawler

- Step 4 :

```
import urllib.request
from bs4 import BeautifulSoup
import re
i = 0
URLdict=dict()#建立dict,分配空间
while i<10:
  if i==0:
    response = urllib.request.urlopen('http://www.entgroup.cn/news/Exclusive/index.shtml')
  else:
    response = urllib.request.urlopen('http://www.entgroup.cn/news/Exclusive/index'+str(i)+'.shtml')
  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())
  i=i+1
```

# *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()
```

# Web Crawler

## *Results*

```
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… | 2016-11-1 |
| 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… | 2016-11-1 |
| 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… | 2016-11-1 |
| 4 | ENAwards最佳娱乐营销案例优秀入围推荐（… | ENAwards最佳娱乐营销案例优秀… | http://www.entgroup.cn/news/Exclusive/18374… | 2016-11-1 |
| 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… | 2016-11-1 |
| 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… | 2016-11-1 |
| 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… | 2016-11-0 |
| 8 | "IP买卖"乱潮中 怎样才能让你的IP卖更多钱 | "IP买卖"乱潮中 怎样才能让你的IP… | http://www.entgroup.cn/news/Exclusive/09373… | 2016-11-0 |
| 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… | 2016-11-1 |
| 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… | 2016-11-1 |
| 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… | 2016-11-1 |
| 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… | 2016-11-1 |
| 13 | 中南文化以收购换业绩成效明显 文化产业… | 中南文化以收购换业绩成效明显 … | http://www.entgroup.cn/news/Exclusive/27371… | 2016-10-2 |
| 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… | 2016-11-1 |
| 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… | 2016-11-0 |
| 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… | 2016-11-0 |
| 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… | 2016-10-2 |
| 18 | ENAwards最佳娱乐营销案例优秀入围推荐（… | ENAwards最佳娱乐营销案例优秀… | http://www.entgroup.cn/news/Exclusive/18374… | 2016-11-1 |

# Web Crawler

***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

```
from flask import Flask
from flask import request
import urllib

app = Flask(__name__)

@app.route('/', methods=['GET', 'POST'])
def home():
    return '''<h1>Home</h1>
              <p><a href="/CityWeather">Weather Forecast</a></p>'''

@app.route('/CityWeather', methods=['GET'])
def login_form():
    return '''<form action="/CityWeather" method="POST">
              <p>City Name <input name="CityName"></p>
              <p><button type="submit">Weather</button></p>
              </form>'''

@app.route('/CityWeather', methods=['POST'])
def login():
    # get data from Baidu:
    url = 'http://apis.baidu.com/heweather/weather/free?city='+request.form['CityName']
    req = urllib.request.Request(url)
    req.add_header("apikey", "5e7ef01f44164283876b0dd0cbd1461e")# Dont forget to change your apikey
    resp = urllib.request.urlopen(req)
    content = resp.read()
    if (content):
        return content
    return '<h3>Bad City Name.</h3>'

if __name__ == '__main__':
    app.run()
```

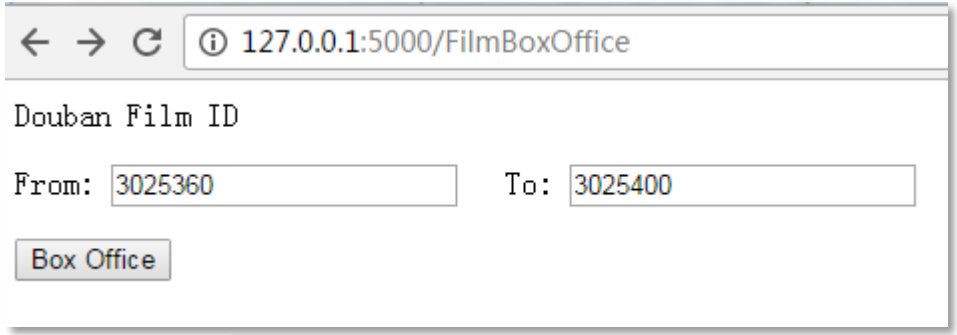
# *API for Douban*

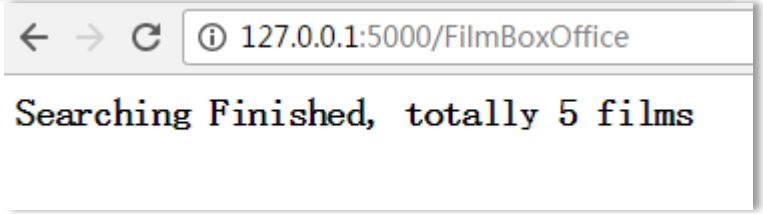
```
# -*- coding: utf-8 -*-
import sys,urllib.request,json

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*

- Set a Range
- Loops
- 1 Douban API
- 2 Saving into DB




```
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 -
```

# *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 Već viđeno（Déjà... | "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


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

# System Integration

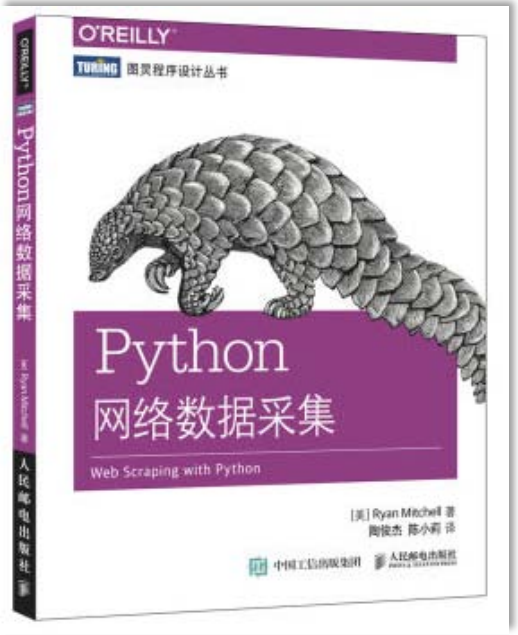
## *Film Box Office Prediction*

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

# Reference

# Reference

## *Python网络数据采集*



# Reference

**Python中使用Beautiful Soup库的超详细教程**

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

# Homework

# Homework

- Try to build a web crawler for your group

# The End of Lecture 7

Thank You

http://www.wangting.ac.cn