



New Media Data Analytics and Application

Lecture 8: Web API Applications

Ting Wang

Outlines

- API Data Storage
- Web Crawlers for APIs
- Web API Applications





tips to the information storage using databases

API Data Storage

API Data Storage

How to save these information into the database?



← → ↺ 127.0.0.1:5000/NewsTouTiao

财经新闻:

Title

- 1 [实体店定价超原产国 加拿大鹅打击代购诚意](#)
- 2 [当U盘已成往事：关于青春，你还记得什么？](#)
- 3 [家庭不和睦，人就会生病！](#)
- 4 [“爸，在忙吗？”这两天，杭州多位家长收到这种短信](#)
- 5 [捐款1个亿！80后夫妇王麒诚和妻子吴艳浙大演讲，满满的创业干货](#)
- 6 [企业出现如下几种现象，仍不进行变革，结局只能是关门大吉](#)
- 7 [东北玉米价位已至高位 或难以持续上涨行情](#)
- 8 [大蒜市场观察：今年大蒜市场遇冷，2019年会上涨吗？](#)
- 9 [这个一线城市二手房市场凉凉，业主降价15%卖房，中介半年没接单](#)
- 10 [龙头点评：目前市场唯一四板，有望领涨大盘（附股）](#)
- 11 [再生水利用将成破解水资源短缺重要途径 千亿级水处理行业迎来新风口](#)
- 12 [四川自贸试验区：为内陆开放添动力](#)
- 13 [国际煤市要闻回顾（11.19-11.23）](#)
- 14 [目前主产区玉米价格仍处于上涨通道中 农民怎么才能卖个最高价](#)
- 15 [家家悦亿元入股福悦祥，11家门店及13万平物流中心尽归麾下！](#)
- 16 [柬埔寨智慧农业领域发展潜力依旧巨大](#)
- 17 [300多天，400余次调控，未来房价如何走](#)
- 18 [实话告诉你！未来10年才是最好的买房时期](#)
- 19 [房价涨愁，房价跌忧，刚需到底怎么了？](#)
- 20 [这6只年报预增翻倍股被机构盯上了！](#)
- 21 [马云讲：没钱该不该创业三句话选择创业新思路](#)
- 22 [股价“梦回2017”，缺钱的复星医药正“骑虎难下”](#)
- 23 [家族信托与集合资金信托的区别是什么？](#)
- 24 [从禁止石油开采到成为全球头号产油国，美国为何突然大量出口石油](#)

Date	Category	Author
2018-11-26 00:29	财经	北京商报
2018-11-26 00:27	财经	中国新闻网
2018-11-26 00:07	财经	一生致富
2018-11-25 23:49	财经	杭州卡洛儿
2018-11-25 23:46	财经	GW34568创业家
2018-11-25 23:24	财经	诺言说职场
2018-11-25 23:23	财经	科学兴农
2018-11-25 23:23	财经	科学兴农
2018-11-25 23:19	财经	每日经济新闻
2018-11-25 23:19	财经	美湖路的老程
2018-11-25 23:17	财经	每日经济新闻
2018-11-25 23:17	财经	新华网
2018-11-25 22:48	财经	懂熊帝
2018-11-25 22:45	财经	每日粮油
2018-11-25 22:41	财经	第一纺织网
2018-11-25 22:40	财经	华彩国际传媒
2018-11-25 22:27	财经	今朝聊房事
2018-11-25 22:27	财经	深度金融
2018-11-25 22:27	财经	天狼说地产
2018-11-25 22:25	财经	21世纪经济报道
2018-11-25 22:24	财经	要树立干事创业的鲜明导向
2018-11-25 22:23	财经	中访网
2018-11-25 22:17	财经	超仁博一博
2018-11-25 22:16	财经	环球财经天下
2018-11-25 22:15	财经	科技财经网



API Data Storage

EXAMPLE 1: Build the News URL List

I ♥ APIs

Get this free sticker at apigee.com



Database Preparation

- Data Extraction from JSON

```
NewsTitle = json.dumps(response_json['data'][i].get("title"), ensure_ascii=False)
NewsDate = json.dumps(response_json['data'][i].get("date"))
NewsCate = json.dumps(response_json['data'][i].get("category"), ensure_ascii=False)
NewsAuthor = json.dumps(response_json['data'][i].get("author_name"), ensure_ascii=False)
NewsURL = json.dumps(response_json['data'][i].get("url"))
```

- Database Columns

- Primary Key
- Columns
- Contents



Database Document

档案名称	JUHE_API_ONLINE_NEWS					
档案用途	聚合数据API新闻原始数据					
主键(PK)	JUHE_API_ONLINE_NEWS_PK: JUHE_API_NEWS_ID(Cluster Index)					
附键(AK)	JUHE_API_ONLINE_NEWS_AK: JUHE_API_NEWS_URL(Unique)					
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	JUHE_API_NEWS_ID	新闻编号	NUMBER		X	
02	JUHE_API_NEWS_TITLE	新闻标题	VARCHAR	300	X	
03	JUHE_API_NEWS_DATE	新闻时间	DATETIME		X	
04	JUHE_API_NEWS_CATE	新闻类别	VARCHAR	100	X	
05	JUHE_API_NEWS_AUTHOR	新闻作者	VARCHAR	100	X	
06	JUHE_API_NEWS_URL	新闻网址	VARCHAR	500	X	
07	JUHE_API_NEWS_CONTENT	新闻内容	TEXT			
08	JUHE_API_NEWS_CONTENT_SEG	内容分词	TEXT			

API Data Storage

SQL Script

```
1 CREATE TABLE JUHE_API_ONLINE_NEWS
2 (
3     JUHE_API_NEWS_ID          INT(10)          PRIMARY KEY AUTO_INCREMENT,
4     JUHE_API_NEWS_TITLE       VARCHAR(300)       NOT NULL,
5     JUHE_API_NEWS_DATE        DATETIME           NOT NULL,
6     JUHE_API_NEWS_CATE        VARCHAR(100)       NOT NULL,
7     JUHE_API_NEWS_AUTHOR      VARCHAR(100)       NOT NULL,
8     JUHE_API_NEWS_URL         VARCHAR(500)       NOT NULL UNIQUE,
9     JUHE_API_NEWS_CONTENT     TEXT,
10    JUHE_API_NEWS_CONTENT_SEG  TEXT
11 );
```




```
response = request1(Newstype, appkey, "GET").replace('\'', '\"').replace('\'\'', '\"').replace('\'xa0', '\"').replace('\'\'', '\"')
response_json = json.loads(response)
```

```
for i in range(30):
    NewsTitle = json.dumps(response_json['data'][i].get("title"), ensure_ascii=False)
    NewsDate = json.dumps(response_json['data'][i].get("date"))
    NewsCate = json.dumps(response_json['data'][i].get("category"), ensure_ascii=False)
    NewsAuthor = json.dumps(response_json['data'][i].get("author_name"), ensure_ascii=False)
    NewsURL = json.dumps(response_json['data'][i].get("url"))
```

to erase unreadable code

```
# 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 JUHE_API_ONLINE_NEWS(JUHE_API_NEWS_TITLE, JUHE_API_NEWS_DATE, JUHE_API_NEWS_CATE, JUHE_API_NEWS_AUTHOR, JUHE_API_NEWS_URL) VALUES('"+NewsTitle[1:-1]+"', '"+NewsDate[1:-1]+"', '"+NewsCate[1:-1]+"', '"+NewsAuthor[1:-1]+"', '"+NewsURL[1:-1]+"')"
```

```
# 4. Cursor Moving
```

```
# 体验游标
```

```
# 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:
```

```
cursor.close()
```

```
# 关闭Connection:
```

```
conn.close()
```

Question: Why not use INSERT INTO?

Answer: The difference between INSERT INTO and REPLACE INTO:

INSERT INTO: INSERT items unconditionally

REPLACE INTO: If not exist, then insert; if exist, then update

When we import data via Juhe Web API frequently, we may download the same data.

Thus we should not save the same data twice in our databases.

Save the data

```
1]+</td></tr>"
ResultTable+="</table>"
# 1. 新闻查询
return CNewstype+"新闻: "+ResultTable
```

Do NOT forget: import pymysql

Juhe News Data
NewsTouTiaoWebFormal.py



上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

API Data Storage

Data saved in the database

JUHE_API_NEWS_ID	JUHE_API_NEWS_TITLE	JUHE_API_NEWS_DATE	JUHE_API_NEWS_CATE	JUHE_API_NEWS_AUTHOR	JUHE_API_NEWS_URL
178	库里KD忍不住想笑！科尔这话基本宣告NBA...	2018-12-09 01:22:00	体育	浣仰	http://mini.eastdav.com/mobile/181209012214377.html
179	上次输球还在17年前！AC米兰打都灵有心得...	2018-12-09 01:21:00	体育	剪影足球	http://mini.eastdav.com/mobile/181209012155740.html
180	欧冠赛制存在改进空间，不合理制度或导致...	2018-12-09 01:21:00	体育	森娜体育	http://mini.eastdav.com/mobile/181209012143092.html
312	领导干部正确对待使用权力，全心全意为人...	2018-12-09 00:21:00	国内	蜜蜂与花的故事	http://mini.eastdav.com/mobile/181209002140799.html
313	飞阅齐鲁 白天看湖畔上观灯，这样的大明...	2018-12-09 00:19:00	国内	齐鲁网	http://mini.eastdav.com/mobile/181209001956179.html
314	孝南：检、企共建，构筑未成年人预防帮教...	2018-12-09 00:14:00	国内	湖北省人民检察院	http://mini.eastdav.com/mobile/181209001445609.html
358	陈奕迅否认新歌抄袭，接受批评的声音，他...	2018-12-09 01:36:00	娱乐	小姐姐的明星梦	http://mini.eastdav.com/mobile/181209013600255.html
359	皇帝——双料影帝，感恩厚待一切归零	2018-12-09 01:33:00	娱乐	韩剧前行	http://mini.eastdav.com/mobile/181209013337667.html
360	斗鱼又一人气主播转型歌手，大司马发布单...	2018-12-09 01:33:00	娱乐	直播吃蛋壳	http://mini.eastdav.com/mobile/181209013321458.html
361	倾城时光：林浅哥哥身份曝光，帮助二家公...	2018-12-09 01:33:00	娱乐	智跑影御	http://mini.eastdav.com/mobile/181209013316682.html
362	同样早薛之谦前女友，嫁对了“武松”的童龄...	2018-12-09 01:32:00	娱乐	昕薇	http://mini.eastdav.com/mobile/181209013238013.html
363	林更新解锁“叛门”新方式，声称吃蛋壳可以...	2018-12-09 01:32:00	娱乐	柯哥玩娱乐	http://mini.eastdav.com/mobile/181209013214645.html
364	李诞将缺席《野生厨房》，自嘲只会搞笑，...	2018-12-09 01:32:00	娱乐	浙纸薇莱声斯	http://mini.eastdav.com/mobile/181209013213675.html
365	“女神”苗苗到底有多俊？就在今天，苗苗亲...	2018-12-09 01:30:00	娱乐	娱乐小站网	http://mini.eastdav.com/mobile/181209013006573.html
366	苗之谦和苗之谦演绎《聪明》，但是雪域的	2018-12-09 01:26:00	娱乐	豪门深闺	http://mini.eastdav.com/mobile/181209012620191.html



Prepare for Web Crawler: Get rid of Flask

- Head: delete flask

```
import json, urllib.request  
from urllib.parse import urlencode  
import pymysql
```

- def request1(): no changes

1

JuheAPIOnlineNewsWebCrawler.py



```
def main():
```

```
# 配置您申请的APPKey
```

```
appkey = "04f*****c17c"
```

```
NewsTypes = ['top', 'shehui', 'guonei', 'guoji', 'yule', 'tiyu', 'junshi', 'keji', 'caijing', 'shishang']
```

```
for NewsType in NewsTypes:
```

```
    if NewsType == "top":
```

```
        CNewsType = "头条"
```

```
    elif NewsType == "shehui":
```

```
        CNewsType = "社会"
```

```
    elif NewsType == "guonei":
```

```
        CNewsType = "国内"
```

```
    elif NewsType == "guoji":
```

```
        CNewsType = "国际"
```

```
    elif NewsType == "yule":
```

```
        CNewsType = "娱乐"
```

```
    elif NewsType == "tiyu":
```

```
        CNewsType = "体育"
```

```
    elif NewsType == "junshi":
```

```
        CNewsType = "军事"
```

```
    elif NewsType == "keji":
```

```
        CNewsType = "科技"
```

```
    elif NewsType == "caijing":
```

```
        CNewsType = "财经"
```

```
    elif NewsType == "shishang":
```

```
        CNewsType = "时尚"
```

```
    else:
```

```
        CNewsType = "头条"
```

1. login() → main()

2. select → list: NewsTypes

2



```

response = request1(NewsType, appkey, "GET").replace('\n', '\n').replace('\n', '\n').replace('\xa0', ',').replace('\n', '\n')
response_json = json.loads(response)
for i in range(30):
    NewsTitle = json.dumps(response_json['data'][i].get("title"), ensure_ascii=False)
    NewsDate = json.dumps(response_json['data'][i].get("date"))
    NewsCate = json.dumps(response_json['data'][i].get("category"), ensure_ascii=False)
    NewsAuthor = json.dumps(response_json['data'][i].get("author_name"), ensure_ascii=False)
    NewsURL = json.dumps(response_json['data'][i].get("url"))

    # 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 JUHE_API_ONLINE_NEWS(JUHE_API_NEWS_TITLE, JUHE_API_NEWS_DATE, JUHE_API_NEWS_CATE, JUHE_API_NEWS_AUTHOR, JUHE_API_NEWS_URL)
VALUES('"+NewsTitle[1:-1]+"','"+NewsDate[1:-1]+"','"+NewsCate[1:-1]+"','"+NewsAuthor[1:-1]+"','"+NewsURL[1:-1]+"')"
    # 4. Cursor Moving
    # 游标执行, 游标移至当前位置
    cursor.execute(sqlstr)
    # 提交事务:
    conn.commit()
    # 5. Connection Close
    # 关闭Cursor:
    cursor.close()
    # 关闭Connection:
    conn.close()

    print(CNewsType + "新闻: " + NewsTitle + " " + NewsDate + " " + NewsURL)
# 1. 新闻查询
print(CNewsType+"新闻抓取完毕! //////////////////////////////////////")

```

3. Delete HTML code, change "return" to "print"

3



API Data Storage



Ask A Question



How to make the program automatically collect data from the API?

Tips:

Loop the “main()”
while(True)





build a web crawler to download contents from urls derived from APIs

Web Crawlers for APIs

API Data Storage

URL List is in the database

JUHE_API_NEWS_ID	JUHE_API_NEWS_TITLE	JUHE_API_NEWS_DATE	JUHE_API_NEWS_CATE	JUHE_API_NEWS_AUTHOR	JUHE_API_NEWS_URL
178	库里KD忍不住想笑！科尔这话基本宣告NBA...	2018-12-09 01:22:00	体育	浣仰	http://mini.eastdav.com/mobile/181209012214377.html
179	上次输球还在17年前！AC米兰打都灵有心得...	2018-12-09 01:21:00	体育	剪影足球	http://mini.eastdav.com/mobile/181209012155740.html
180	欧冠赛制存在改进空间，不合理制度或导致...	2018-12-09 01:21:00	体育	森娜体育	http://mini.eastdav.com/mobile/181209012143092.html
312	领导干部正确对待使用权力，全心全意为人...	2018-12-09 00:21:00	国内	蜜蜂与花的故事	http://mini.eastdav.com/mobile/181209002140799.html
313	飞阅齐鲁 白天看湖畔上观灯，这样的大明...	2018-12-09 00:19:00	国内	齐鲁网	http://mini.eastdav.com/mobile/181209001956179.html
314	孝南：检、企共建，构筑未成年人预防帮教...	2018-12-09 00:14:00	国内	湖北省人民檢察院	http://mini.eastdav.com/mobile/181209001445609.html
358	陈奕迅否认新歌抄袭，接受批评的声音，他...	2018-12-09 01:36:00	娱乐	小姐姐的明星梦	http://mini.eastdav.com/mobile/181209013600255.html
359	早帝——双料影帝，感恩厚待一切归零	2018-12-09 01:33:00	娱乐	鲜影前行	http://mini.eastdav.com/mobile/181209013337667.html
360	斗鱼又一人气主播转型歌手，大司马发布单...	2018-12-09 01:33:00	娱乐	直播吃蛋壳	http://mini.eastdav.com/mobile/181209013321458.html
361	倾城时光：林浅哥哥身份曝光，帮助三家公...	2018-12-09 01:33:00	娱乐	智跑影御	http://mini.eastdav.com/mobile/181209013316682.html
362	同样早薛之谦前女友，嫁对了“武松”的童玲...	2018-12-09 01:32:00	娱乐	昕薇	http://mini.eastdav.com/mobile/181209013238013.html
363	林更新解锁“扳门”新方式，声称吃蛋壳可以...	2018-12-09 01:32:00	娱乐	柯哥玩娱乐	http://mini.eastdav.com/mobile/181209013214645.html
364	李诞将缺席《野生厨房》，自嘲只会搞笑，...	2018-12-09 01:32:00	娱乐	浙纸壹壹声嘶	http://mini.eastdav.com/mobile/181209013213675.html
365	“女神”黄渤到底有多傻？就在今天，黄渤亲...	2018-12-09 01:30:00	娱乐	娱乐小站网	http://mini.eastdav.com/mobile/181209013006573.html
366	薛之谦和杨幂演绎《聪明》，但是雪域的	2018-12-09 01:26:00	娱乐	豪门深闺	http://mini.eastdav.com/mobile/181209012620191.html

How to get the content by these URLs?



Analyze the Structure of the Potential Candidate Website

- Employ a good web explorer
Eg. Google Chrome
- Select a webpage as a start,
or prepare a list for web crawler.



Web Crawler

EXAMPLE 2: News Web Crawler

Web Crawler



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
 - 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>



A Review:

How to collect data from the Website of SHISU?

```
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)
```

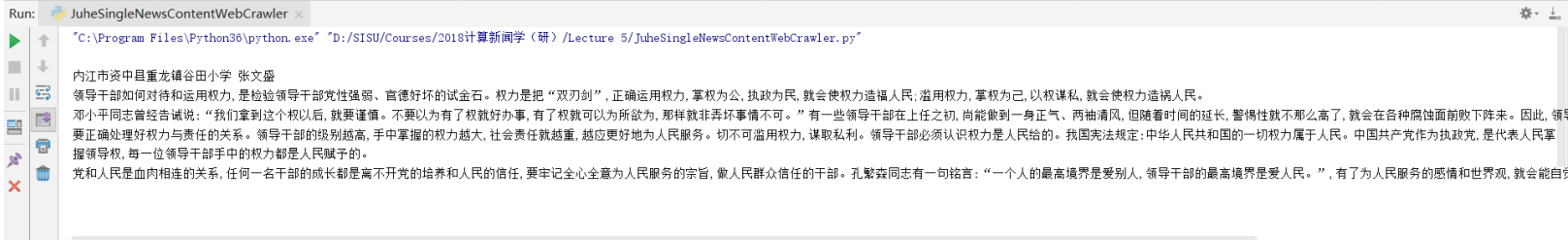


Revise the code for Juhe News API

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

*Modular Construction: Blocks on blocks.
Integrate this code with the Juhe API Code.*

```
response = urllib.request.urlopen('http://mini.eastday.com/mobile/181209002140799.html')  
HTMLText = response.read()  
BSobj = BeautifulSoup(HTMLText,"html.parser")#基于BeautifulSoup分析整个页面  
  
ArticleContent = BSobj.find("div", {"class": "J-article-content article-content"})  
print(ArticleContent.get_text())
```





领导干部正确对待使用权力，全心全意为人民服务

div#content.J-article-content.article-content | 718×1387.3

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。

党和人民是血肉相连的关系,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



Elements Console >> 2

> <div id="title">...</div>

... > <div id="content" class="J-article-content article-content" data-pswp-uid="1">...</div> == \$0

</article>

> <div class="articledown-wrap gg-item J-gg-item" data-ggpos="articledown" data-pgnum="1" data-tiidx="-3"> </div>

... #/J_article div#content.J-article-content.article-content

Styles Computed Event Listeners DOM Breakpoints >>

Filter :hov .cls +

element.style {

}

#content { page_details_v5.min.css:1

margin: > .3rem 16px 16px;

text-align: justify;

}

article, aside, blockquote, body, common.min.css:1

button, code, dd, details, div, dl, dt, fieldset,

figcaption, figure, footer, form, h1, h2, h3, h4, h5,

h6, header, hgroup, hr, input, legend, li, menu, nav,

: Console What's New X

Highlights from the Chrome 70 update

Live Expressions in the Console

Pin expressions to the top of the Console to monitor their values in real-time.

Part 1: Import and Get_text()

```
1 import json, urllib.request
2 from urllib.parse import urlencode
3 import pymysql
4 from bs4 import BeautifulSoup
5
6 def NewsArticleContent(WebURL):
7     response = urllib.request.urlopen(WebURL)
8     HTMLText = response.read()
9     BSobj = BeautifulSoup(HTMLText, "html.parser") #基于BeautifulSoup分析整个页面
10
11     ArticleContent = BSobj.find("div", {"class": "J-article-content article-content"})
12     return ArticleContent.get_text()
```



JuheAPIOnlineNewsWebCrawlerWithContent.py



Part 2: get news content in main()

```
44 response = requestl(NewsType, appkey, "GET").replace('\n', '\n').replace('\r', '\r').replace('\xa0', ' ').replace('\u2013', '-')
45 response_json = json.loads(response)
46 for i in range(30):
47     NewsTitle = json.dumps(response_json['data'][i].get("title"), ensure_ascii=False)
48     NewsDate = json.dumps(response_json['data'][i].get("date"))
49     NewsCate = json.dumps(response_json['data'][i].get("category"), ensure_ascii=False)
50     NewsAuthor = json.dumps(response_json['data'][i].get("author_name"), ensure_ascii=False)
51     NewsURL = json.dumps(response_json['data'][i].get("url"))
52
53     NewsContent = NewsArticleContent(NewsURL[1:-1]).replace('\n', '\n').replace('\r', '\r')
54
55     # save data in database:
56
57     # 连接到MySQL数据库
58     # 1.Connection Open
59     conn = pymysql.connect(user='root', password='123456', database='raw_data', charset='utf8')
```



Part 3: insert into database in main()

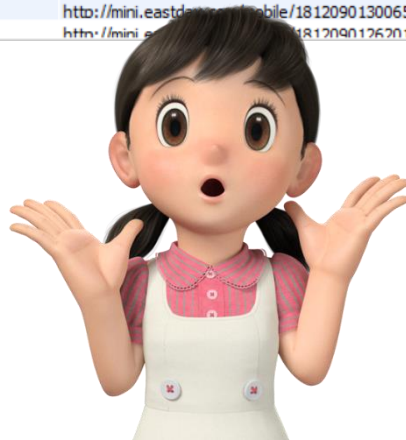
```
62  
63  
64 _URL, JUHE_API_NEWS_CONTENT) VALUES(''+NewsTitle[1:-1]+'', ''+NewsDate[1:-1]+'', ''+NewsCate[1:-1]+'', ''+NewsAuthor[1:-1]+'', ''+NewsURL[1:-1]+'', ''+NewsContent+'')  
65  
66
```



Web Crawler

Results

NEWS_ID	JUHE_API_NEWS_TITLE	JUHE_API_NEWS_DATE	JUHE_API_NEWS_CATE	JUHE_API_NEWS_AUTHOR	JUHE_API_NEWS_URL	JUHE_API_NEWS_CONTENT
	匡甲KD忍不住想笑！科尔这话基本宣告NBA...	2018-12-09 01:22:00	体育	浣仰	http://mini.eastdav.com/mobile/181209012214377.html	在杜兰特表现低迷的情况下，勇士今天还...
	上次输球还在17年前！AC米兰打都灵有心得...	2018-12-09 01:21:00	体育	剪影足球	http://mini.eastdav.com/mobile/181209012155740.html	北京时间12月10日凌晨3点30分，AC米兰将...
	欧冠赛制存在改进空间，不合理制度或导致...	2018-12-09 01:21:00	体育	森娜体育	http://mini.eastdav.com/mobile/181209012143092.html	我想在这里提出一个太阳的设想，欧冠联...
	领导干部正确对待使用权力，全心全意为人...	2018-12-09 00:21:00	国内	蜜蜂与花的故事	http://mini.eastdav.com/mobile/181209002140799.html	内江市资中县重龙镇谷田小学 张文盛 领导...
	飞阅齐鲁！白天看湖畔上观灯，这样的太阳...	2018-12-09 00:19:00	国内	齐鲁网	http://mini.eastdav.com/mobile/181209001956179.html	齐鲁网12月8日讯空中纵览山东，品味时代...
	孝南：检、企共建，构筑未成年人预防犯罪...	2018-12-09 00:14:00	国内	湖北省人民检察院	http://mini.eastdav.com/mobile/181209001445609.html	本网讯（通讯员 董耕霞）孝南区未成年...
	陈奕迅否认新歌抄袭，接受批评的声音，他...	2018-12-09 01:36:00	娱乐	小胡胡的明星梦	http://mini.eastdav.com/mobile/181209013600255.html	华语乐坛有很多杰出的歌手，如周杰伦、...
	皇帝——双料影帝，感恩厚待一切归零	2018-12-09 01:33:00	娱乐	魅影前行	http://mini.eastdav.com/mobile/181209013337667.html	皇帝导演荣获“第十届中国电影金鸡奖”...
	斗鱼又一人气主播转型歌手，大司马发布单...	2018-12-09 01:33:00	娱乐	直播吃蛋卷	http://mini.eastdav.com/mobile/181209013321458.html	斗鱼有着这样一位主播，凭着幽默风趣的...
	倾城时光：林浅哥哥身份曝光，帮助二家公...	2018-12-09 01:33:00	娱乐	智跑影视	http://mini.eastdav.com/mobile/181209013316682.html	赵丽颖主演的电视剧《你和我的倾城时光...
	同样是薛之谦前女友，嫁对了“武松”的童玲...	2018-12-09 01:32:00	娱乐	昕薇	http://mini.eastdav.com/mobile/181209013238013.html	《亲爱的客栈2》热度不断，继“明星”美...
	林更新解锁“扳门”新方式，声称吃虾壳可以...	2018-12-09 01:32:00	娱乐	柯哥玩娱乐	http://mini.eastdav.com/mobile/181209013214645.html	林更新解锁“扳门”新方式，声称吃虾壳可以...
	李诞将缺席《野生厨房》，白敬明会搞笑，...	2018-12-09 01:32:00	娱乐	浙纸壹壹声嘶	http://mini.eastdav.com/mobile/181209013213675.html	《野生厨房》是由汪涵主持的一档节目，...
	“女神”董洁到底有多傻？就在今天，董洁亲...	2018-12-09 01:30:00	娱乐	娱乐小站风	http://mini.eastdav.com/mobile/181209013006573.html	微博热搜每天变换不断，成为当今大多数...
	薛之谦指挥空袭警报《啦啦》，但最震撼的...	2018-12-09 01:26:00	娱乐	高门深巷	http://mini.eastdav.com/mobile/181209012620191.html	第十二届音乐盛典咪咕汇正在火热进行中



Tips:

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





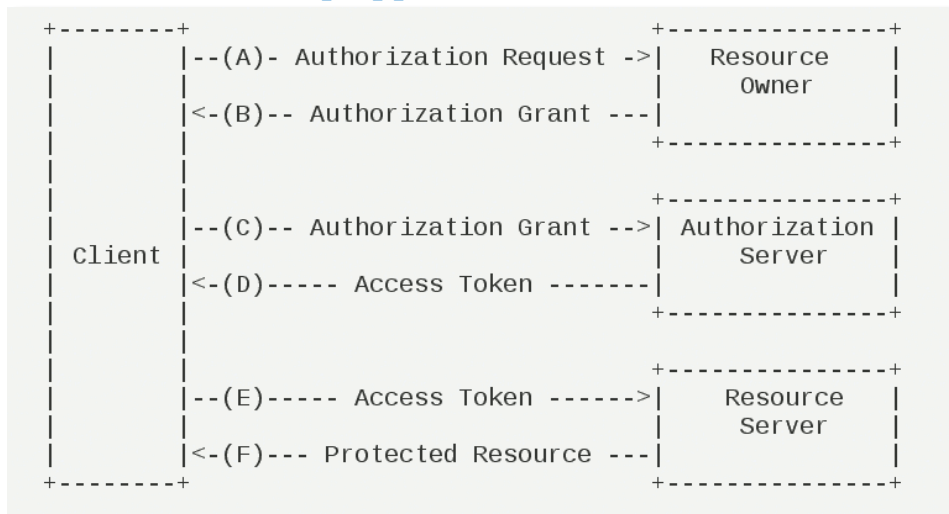
some examples on social media

Web API Applications

Web API Applications

OAuth

An open protocol to allow secure API authorization in a simple and standard method from web, mobile and desktop applications.



API Key
needed

Access Token
needed



Ref. http://www.ruanyifeng.com/blog/2014/05/oauth_2_0.html



Web API Applications

EXAMPLE 3: Friendship of TVs on Twitter



Friendship of TVs on Twitter

Problem Description

This research aims to investigate the significance and activeness of TV media companies in the world. Twitter friendships between TV broadcasting companies were employed for this research. Totally, about 104 TV channels from different countries are selected for this research. In-Degree and Out-Degree are used as features in this study.

Features:

In-Degree: Followed by Others, which shows the significance to the world

Out-Degree: Following Others, which shows the activeness in the world



Friendship of TVs on Twitter

twitter and Its API

twitter

Do NOT tell me you do NOT know twitter, even you are a Chinese.

If you really do not know it, visit <https://twitter.com> right now!



twitter API: <https://dev.twitter.com/>



Friendship of TVs on Twitter

Tweepy

An easy-to-use Python library for accessing the Twitter API

Installation

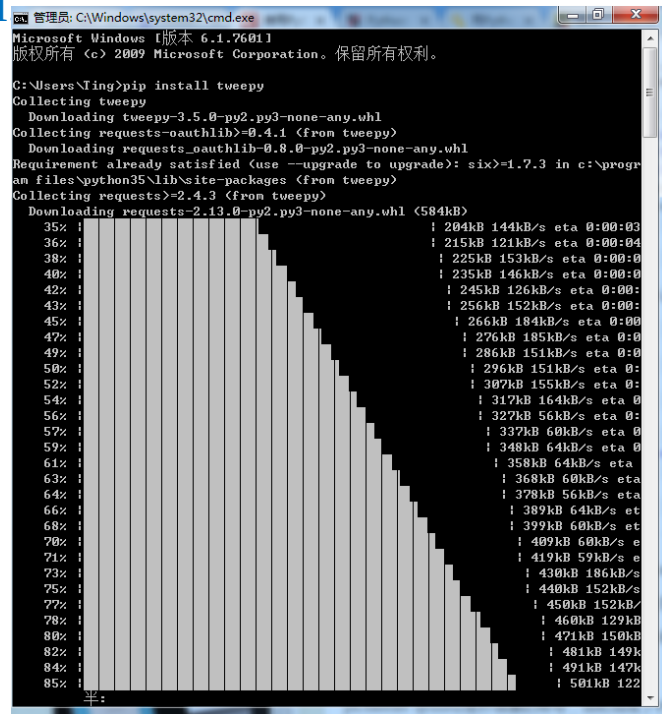
pip install tweepy

Reference to Tweepy

<https://github.com/tweepy/tweepy>

<http://www.tweepy.org/>

<http://pythonhosted.org/tweepy/>



```
管理员: C:\Windows\system32\cmd.exe
Microsoft Windows [版本 6.1.7601]
版权所有 (c) 2009 Microsoft Corporation. 保留所有权利。

C:\Users\Ting>pip install tweepy
Collecting tweepy
  Downloading tweepy-3.5.0-py2.py3-none-any.whl
Collecting requests-oauthlib-0.4.1 (from tweepy)
  Downloading requests_oauthlib-0.8.0-py2.py3-none-any.whl
Requirement already satisfied (use --upgrade to upgrade): six>=1.7.3 in c:\program files\python35\lib\site-packages (from tweepy)
Collecting requests>=2.4.3 (from tweepy)
  Downloading requests-2.13.0-py2.py3-none-any.whl (584kB)
35% | 204kB 144kB/s eta 0:00:03
36% | 215kB 121kB/s eta 0:00:04
38% | 225kB 153kB/s eta 0:00:0
40% | 235kB 146kB/s eta 0:00:0
42% | 245kB 126kB/s eta 0:00:0
43% | 256kB 152kB/s eta 0:00:0
45% | 266kB 184kB/s eta 0:00:0
47% | 276kB 185kB/s eta 0:00:0
49% | 286kB 151kB/s eta 0:00:0
50% | 296kB 151kB/s eta 0:00:0
52% | 307kB 155kB/s eta 0:00:0
54% | 317kB 164kB/s eta 0:00:0
56% | 327kB 56kB/s eta 0:00:0
57% | 337kB 60kB/s eta 0:00:0
59% | 348kB 64kB/s eta 0:00:0
61% | 358kB 64kB/s eta 0:00:0
63% | 368kB 60kB/s eta 0:00:0
64% | 378kB 56kB/s eta 0:00:0
66% | 389kB 64kB/s eta 0:00:0
68% | 399kB 60kB/s eta 0:00:0
70% | 409kB 60kB/s eta 0:00:0
71% | 419kB 59kB/s eta 0:00:0
73% | 430kB 186kB/s eta 0:00:0
75% | 440kB 152kB/s eta 0:00:0
77% | 450kB 152kB/s eta 0:00:0
78% | 460kB 129kB/s eta 0:00:0
80% | 471kB 150kB/s eta 0:00:0
82% | 481kB 149kB/s eta 0:00:0
84% | 491kB 147kB/s eta 0:00:0
85% | 501kB 122kB/s eta 0:00:0
```



Friendship of TVs on Twitter

Register Your Twitter APPs

Step 1 : visit <https://apps.twitter.com/>, and click on the button “Create New APP”.

 Application Management

Twitter Apps

You don't currently have any Twitter Apps.

Create New App



Friendship of TVs on Twitter

Step2 : Input the detailed information of your new app.

Create an application

Application Details

Name *

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.

(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

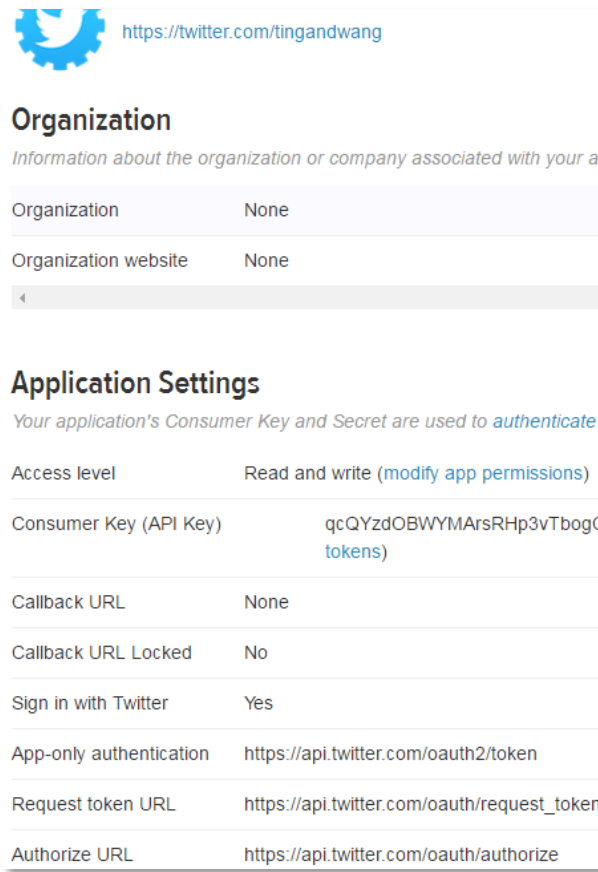
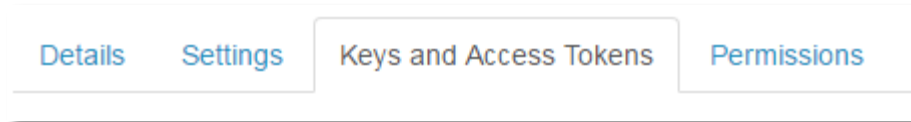


Friendship of TVs on Twitter

Step3: Get your API keys and Access Tokens

[Notes]

You may get all these information when you click on **“Keys and Access Tokens”**.



Friendship of TVs on Twitter

Steps for the System

1. Connect to twitter by API Keys and Access Tokens
2. Find the relationship using Tweepy functions

```
API.show_friendship(source_id/source_screen_name, target_id/target_screen_name)
```

Returns detailed information about the relationship between two users.

- Parameters:
- **source_id** – The user_id of the subject user.
 - **source_screen_name** – The screen_name of the subject user.
 - **target_id** – The user_id of the target user.
 - **target_screen_name** – The screen_name of the target user.

Return type: `Friendship` object

3. Save the results



Friendship of TVs on Twitter

Results

In-Degree for the significance of TV media

In-Degree Ranking

	Country	TV Media	Twitter ID	In-Degree
1	USA	CNN	CNN	26
2	Russia	Russia Today	RT_com	14
3	UK	BBC	BBCNews	13
4	Qatar	Al Jazeera television	AJENews	6
5	Spain	Radio Televisión Española	rtve	5
6	Netherland	Nederlandse Omroep Stichting	NOS	2
7	South Africa	South African Broadcasting Corporation, SABC	SABCNewsOnline	2
8	Lebanon	Al-Manar TV	almanarnews	2



Friendship of TVs on Twitter

Out-Degree for the activeness of TV media

Out-Degree Ranking

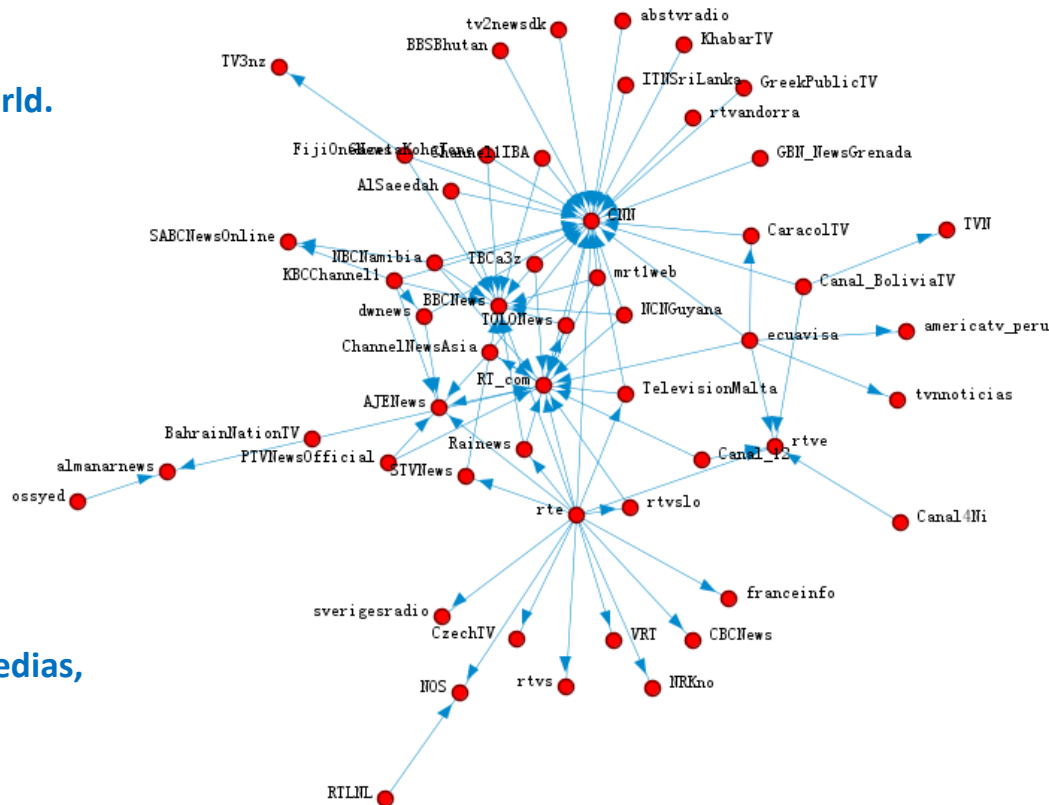
	Country	TV Media	Twitter ID	Out-Degree
1	Ireland	Raidió Teilifís Éireann, Ireland's National Public Service Multi-Media Organisation	rte	17
2	Ecuador	Ecuavisa	ecuavisa	6
3	Kenya	Kenya Broadcasting Corporation	KBCChannel1	5
4	Namibia	Namibian Broadcasting Corporation	NBCNamibia	4
5	Russia	Russia Today	RT_com	4
6	Bolivia	Bolivia TV	Canal_BoliviaTV	3
7	Singapore	Television Corporation of Singapore	ChannelNewsAsia	3
8	Fiji	Fiji Television Limited	FijiOneNews	3
9	Guyana	National Communications Network, Guyana	NCNGuyana	3
10	Tonga	Tonga Broadcasting Commission	TBCa3z	3
11	Afghanistan	TOLONews	TOLONews	3
12	Macedonia	Macedonian Radio-Television	mrt1web	3



Friendship of TVs on Twitter

Conclusions

1. CNN, BBCNews, and Russia Today have the greatest significance to the world.
2. Raidió Teilifís Éireann from Ireland is the most active TV media.
3. Media from developed countries have greater influence.
4. Media from developing countries are more active.
5. CCTV, China neither follow any other medias, nor be followed by any other medias.



Web API Applications

An Extension of Friendship of TVs on Twitter

The same technology can be extended to:

Friendship between

companies,
governments,
organizations,
universities,
people...





上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Reference

Reference

Baidu	(http://apistore.baidu.com/)
Weibo	(http://open.weibo.com/wiki/微博API)
WeChat	(http://dev.wechat.com/)
QQ Tencent	(http://wiki.open.qq.com/wiki/API文档)
360	(http://open.360.cn/)
Taobao	(http://open.taobao.com/docs/api_list.htm?spm=a219a.7386653.1.22.3XIaLs)
JD	(http://jos.jd.com/)
Suning	(http://open.suning.com/)
Douban	(https://developers.douban.com/)
Juhe	(https://www.juhe.cn/)
Google	(https://developers.google.com/)
Facebook	(http://developers.facebook.com/)
Twitter	(https://dev.twitter.com/)
Linkedin	(https://developer.linkedin.com/)



Social Media Mining

— <http://dmml.asu.edu/smm/>

Social Media Mining

Social Media Mining

Home Download Book Slides/Tutorials Table of Contents Errata How to Cite

Social Media Mining

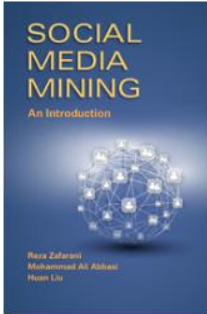
An Introduction

A Textbook by Cambridge University Press

Reza Zafarani
Mohammad Ali Abbasi
Huan Liu

Syracuse University
Machine Zone
Arizona State University

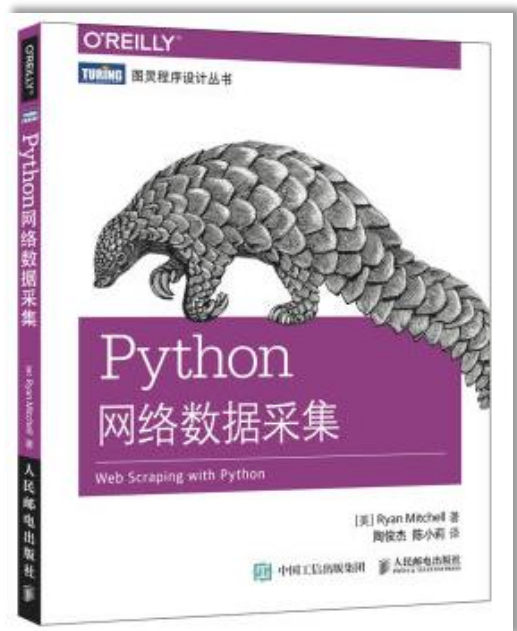
PDF  DOWNLOAD 



Accessed 90,000+ times
from 160+ countries and 1200+ Universities

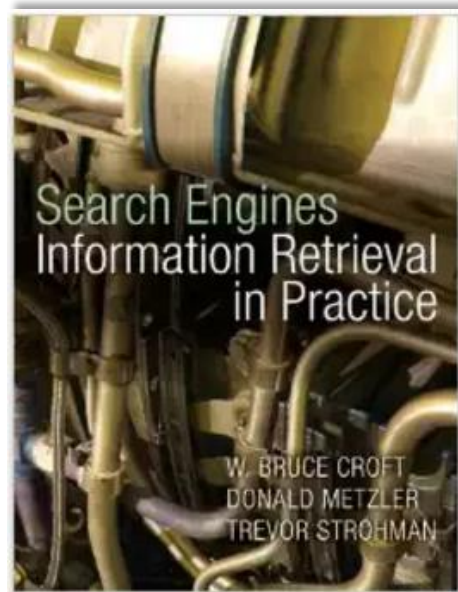
Python网络数据采集

- <https://item.jd.com/11896401.html>
- <http://download.csdn.net/detail/u010309742/9647121?web=web>



search engine information retrieval in practice

- <http://www.search-engines-book.com/>
- <http://www.amazon.com/Search-Engines-Information-Retrieval-Practice/dp/0136072240>



Python中使用Beautiful Soup库的超详细教程

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





The End of Lecture 8

Thank You

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

