上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Computational Journalism

Lecture 5: Information Acquisition using Web Crawlers

Ting Wang

Outlines

- API Data Storage
- Product-Oriented Data Collection
- Web Crawler

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Save the data obtained by API

API Data Storage

API Data Storage

EXAMPLE 1:
Build the News URL List

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			

SQL Script

```sql
CREATE TABLE JUHE_API_ONLINE_NEWS
(
 JUHE_API_NEWS_ID            INT(10)          PRIMARY KEY AUTO_INCREMENT,
 JUHE_API_NEWS_TITLE         VARCHAR(300)     NOT NULL,
 JUHE_API_NEWS_DATE          DATETIME         NOT NULL,
 JUHE_API_NEWS_CATE          VARCHAR(100)     NOT NULL,
 JUHE_API_NEWS_AUTHOR        VARCHAR(100)     NOT NULL,
 JUHE_API_NEWS_URL           VARCHAR(500)     NOT NULL UNIQUE,
 JUHE_API_NEWS_CONTENT       TEXT,
 JUHE_API_NEWS_CONTENT_SEG   TEXT
);
```

```
    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"))

        # 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()

        ResultTable+="<tr><td>"+str((i+1))+"</td><td><a href='"+NewsURL[1:-1]+"' target='_blank'>"+NewsTitle[1:-1]+"</a></td><td>"+NewsDate[1:-1]+"</td><td>"+NewsCate[1:-1]+"</td><td>"+NewsAuthor[1:-1]+"</td></tr>"
    ResultTable+="</table>"
    # 1.新闻查询
    return CNewsType+"新闻："+ResultTable
```

Save the data

Do NOT forget: import pymysql

Juhe News Data

NewsTouTiaoWebFormal.py

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！科尔这话基本盲告NBA...	2018-12-09 01:22:00	体育	魄报	http://mini.eastday.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	直播足球	http://mini.eastday.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.eastday.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.eastday.com/mobile/181209002140799.html
313	飞阅齐鲁丨白天看湖畔上观灯，这样的太明...	2018-12-09 00:19:00	国内	齐鲁网	http://mini.eastday.com/mobile/181209001956179.html
314	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.eastday.com/mobile/181209001445609.html
358	陈奕讯否认新歌抄袭，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.eastday.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	鲜影前行	http://mini.eastday.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.eastday.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	智聊影视	http://mini.eastday.com/mobile/181209013316682.html
362	同样是薛之谦前女友，嫁对了武松的意龄...	2018-12-09 01:32:00	娱乐	昕薇	http://mini.eastday.com/mobile/181209013238013.html
363	林更新解锁"抠门"新方式，吉称吃虾壳可以...	2018-12-09 01:32:00	娱乐	柯哥玩娱乐	http://mini.eastday.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	浙纸爸兼吉斯	http://mini.eastday.com/mobile/181209013213675.html
365	"女版黄渤"到底有多像？就在今天，黄渤亲...	2018-12-09 01:30:00	娱乐	娱乐小钻风	http://mini.eastday.com/mobile/181209013006573.html
366	薛之谦现场深情演绎《路过》，但是雷撒的	2018-12-09 01:26:00	娱乐	亮门逐鹿	http://mini.eastday.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('\"', '"').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"))

    # 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”

API Data Storage

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

Tips:

Loop the “main()”

while(True)

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

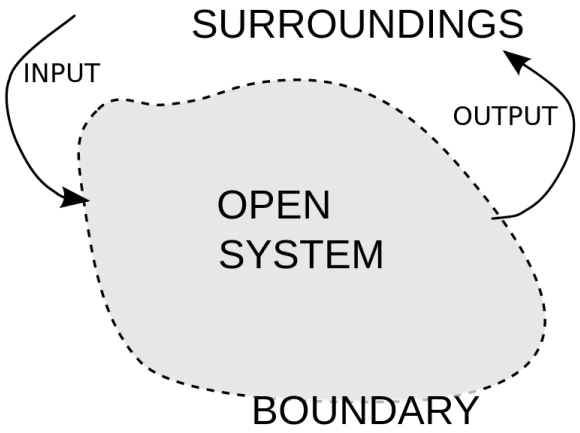
You should know your purpose of data collection

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

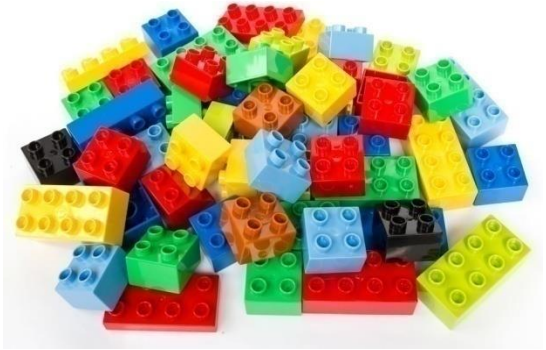
Thirdly, different data should be fused and stored together, just as a database for a product

- 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

a brief introduction to

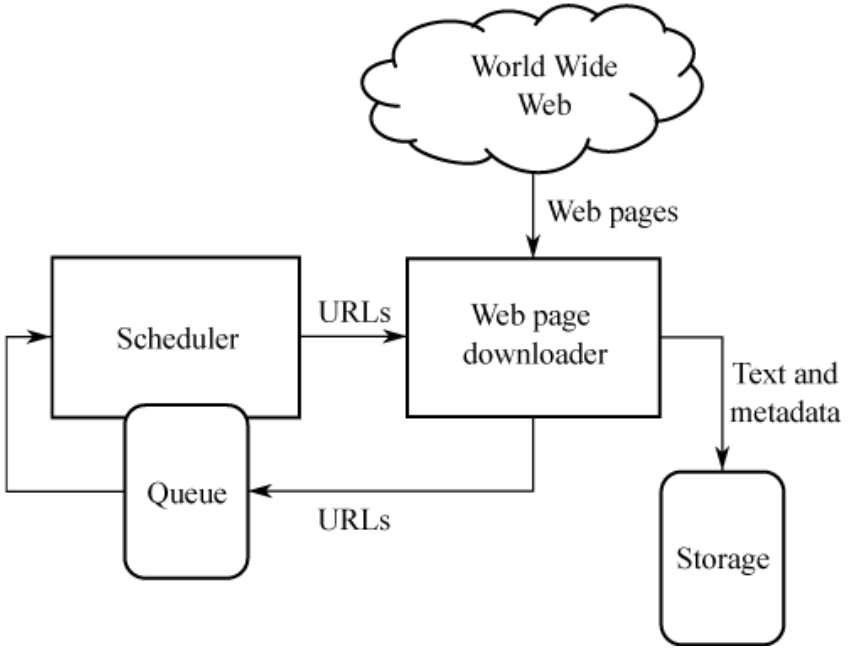
Web Crawler

Web Crawler 网络爬虫

an Internet robot which systematically browses the World Wide Web

Also know as:

- Web Search Engine
- Web Spider
- Web Crawling Robot



Web Crawler

How to design a web crawler?

- Loops: While, for …
- Re-visit: If … elif…else

Four Important Crawling Policies

- **selection policy** states the pages to download
- **re-visit policy** states when to check for changes to the pages
- **politeness policy** states how to avoid overloading Web sites
- **parallelization policy** states how to coordinate distributed web crawlers

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,
  or prepare a list for web crawler.

Web Crawler

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！科尔这话基本盲告NBA...	2018-12-09 01:22:00	体育	魏邱	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	鹏影足球	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	飞阅齐鲁丨白天看湖畔上观灯，这样的太明...	2018-12-09 00:19:00	国内	齐鲁网	http://mini.eastdav.com/mobile/181209001956179.html
314	老南：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
358	陈奕讯否认新歌抄袭，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	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	智聊影视	http://mini.eastdav.com/mobile/181209013316682.html
362	同样是薛之谦前女友，嫁对了"武松"的意龄...	2018-12-09 01:32:00	娱乐	昕薇	http://mini.eastdav.com/mobile/181209013238013.html
363	林更新解锁"抠门"新方式，吉称吃虾壳可以...	2018-12-09 01:32:00	娱乐	柯哥玩娱乐	http://mini.eastdav.com/mobile/181209013214645.html
364	李诞将缺席《野生厨房》，自嘲只会搞笑，...	2018-12-09 01:32:00	娱乐	浙纸爸兼吉斯	http://mini.eastdav.com/mobile/181209013213675.html
365	"女版黄渤"到底有多傻？就在今天，黄渤亲...	2018-12-09 01:30:00	娱乐	娱乐小钻风	http://mini.eastdav.com/mobile/181209013006573.html
366	薛之谦现场深情演绎《路过》，但是雷撒的	2018-12-09 01:26:00	娱乐	高门逐鹿	http://mini.eastdav.com/mobile/181209012620191.html

How to get the content by these URLs?

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

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す(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
```

Web Crawler

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)
```

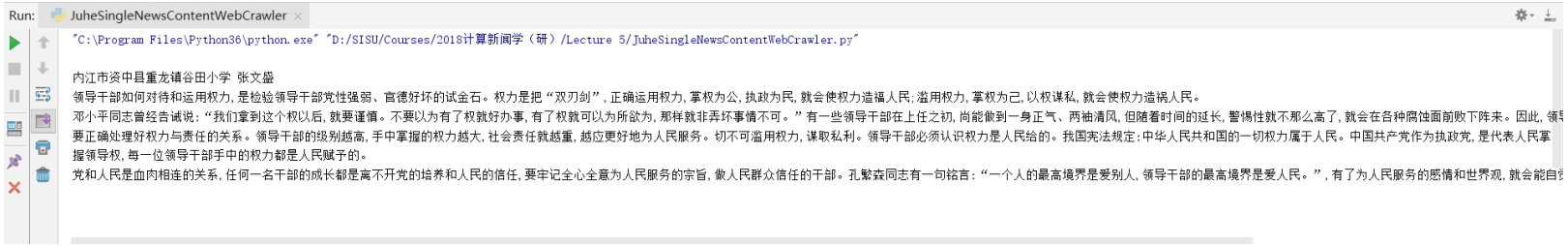
Revise the code for Juhe News API

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

```
import urllib.request
from bs4 import BeautifulSoup

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,就会使权力造祸人民。

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。

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 = request1(NewsType, appkey, "GET").replace('\"', '"').replace('\'', '\"').replace('\xa0',',').replace('\\',',')
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("\'","'").replace("\"",'"')
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

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

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
	库里KD忍不住想笑！科尔谈话基本盲告NBA...	2018-12-09 01:22:00	体育	澎湃	http://mini.eastdav.com/mobile/181209012214377.html	在杜兰特表现低迷的情况下，重十今天还...
	上次输球还在17年前！AC米兰打都录有心得...	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大胆的设想，欧冠联...
	新昌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用小学 张文盛 新昌...
	飞阅齐鲁丨白天看湖畔上观灯，这样的大明...	2018-12-09 00:19:00	国内	齐鲁网	http://mini.eastdav.com/mobile/181209001956179.html	齐鲁网12月8日讯空中纵览山东，品阅时代...
	麦南：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 戴琳霞）麦南区未成年...
	陈奕讯否认新歌抄袭，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	华语乐坛有很多本出的歌手，如周杰伦、...
	导京——双料影帝，感恩厚待 一切归零	2018-12-09 01:33:00	娱乐	魅影循行	http://mini.eastdav.com/mobile/181209013337667.html	导京导演荣获"第十十届华表奖优秀男演...
	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	斗鱼有着这样一位主播，佟着幽默风趣的...
	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	智聊影视	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，嫉对了"武松"的意龄...	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"明天"美...
	林更新解锁"抠门"新方式，吉称吃虾壳可以...	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"抠门"新方式，吉称吃虾壳可以...
	李诞将缺席《野生厨房》，自嘲只会搞笑，...	2018-12-09 01:32:00	娱乐	浙纸黄第吉斯	http://mini.eastdav.com/mobile/181209013213675.html	《野生厨房》是由汪涵主持的一档节目，...
	"女版黄渤"到底有多像？就在今天，黄渤亲...	2018-12-09 01:30:00	娱乐	娱乐小钻风	http://mini.eastdav.com/mobile/181209013006573.html	微博热搜每天变换不断，成为当今大多数...
	薛之谦现场变情演绎《骆驼》，但最震撼的	2018-12-09 01:26:00	娱乐	亭门家鹿	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

Web Crawler

Tips:

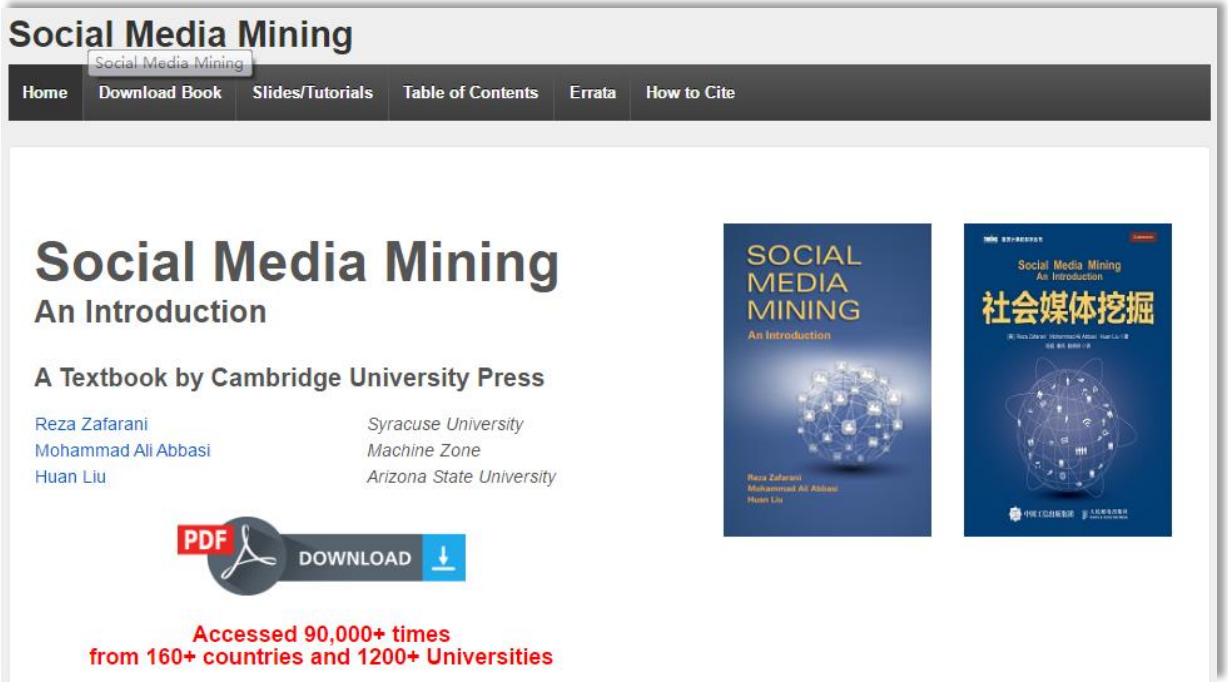
- Web Crawler is very complex. If you want to use web Crawler, you should build them individually for each different websites.

Reference

Reference

Social Media Mining

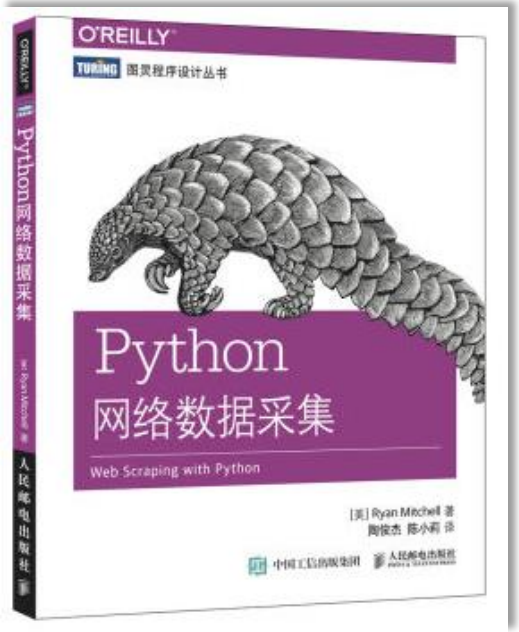
– http://dmml.asu.edu/smm/



Reference

Python网络数据采集

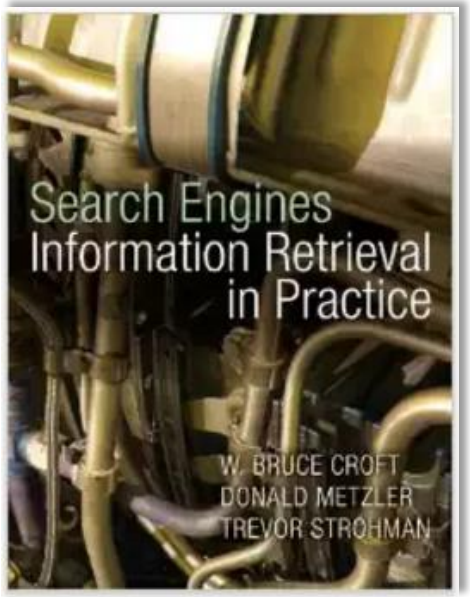
- https://item.jd.com/11896401.html
- http://download.csdn.net/detail/u010309742/9647121?web=web



Reference

search engine information retrieval in practice

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



Reference

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

The End of Lecture 5

Thank You

http://www.wangting.ac.cn