第一次作业：
python网络爬虫介绍

赵阳

北京市海淀区中关村东路95号

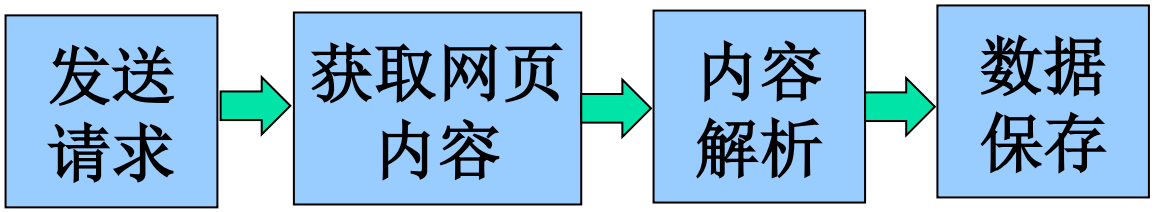
邮件： yang.zhao@nlpr.ia.ac.cn

网络爬虫

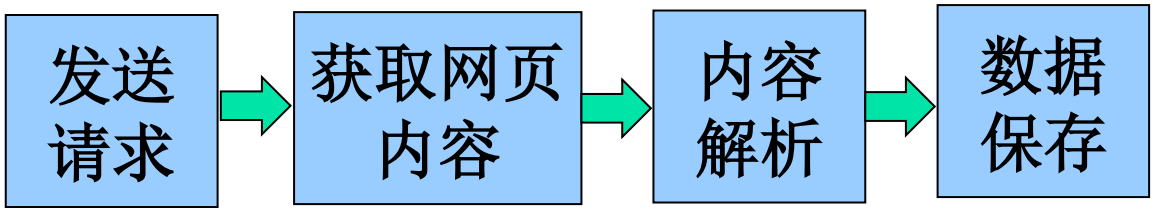
网络爬虫

网络爬虫是一种按照一定的规则，自动地抓取万维网信息的程序

基本过程



网络爬虫



1、发送请求

使用http库向目标站点发起请求。

2、获取网页内容

如果请求内容存在于目标服务器上，那么服务器会返回请求内容。

返回内容包含：html、图片和视频等。

3、内容解析

解析html数据，利用正则表达式或者其他库提取目标信息。

第三方解析库如Beautifulsoup等

4、数据保存

解析得到的数据保存在本地。

软件和库介绍

■ 预备知识

Python的基本语法和编程基础

■ 额外的库

Requests库

(pip install requests)

BeautifulSoup库

(pip install beautifulsoup4)

■ 初学者建议

使用**python**的集成开发工具 **pycharm**

https://www.jetbrains.com/pycharm/

基本框架

```
# coding=utf-8
import requests  ##导入requests库

def getHTMLText(url):  ##python函数
    try:
        r = requests.get(url)  ##发送请求
        r.encoding = r.apparent_encoding  ##获取相应内容编码
        return r.text  ##r.text为网页内容
    except:
        print("爬取失败")

if __name__ == '__main__':  ##main函数
    url = 'http://www.news.cn/world/2022-04/07/c_1128537167.htm/'
    print(getHTMLText(url))
```

核心代码

基本框架

假设网页为下：

首页 | 时政 | 国际 | 财经 | 视频 | 富媒体 | 科技 | 文化 | 健康 | 军事 | 思客智库 | 政务 | 商务

学习进行时 高层 全球连线 理论 网评 法治 人事 廉政 地方 港澳 台湾 教育 科普 体育 直播 访谈 图片 信息化 上市公司 乡村振兴 中央文件 +

新华网 > 国际 > 正文

2022
04/07
00:41:39
来源：新华网

王毅同以色列候任总理、外长拉皮德通电话

字体： 小 中 大 分享到：

新华社北京4月6日电 4月6日，国务委员兼外长王毅应约同以色列候任总理、外长拉皮德通电话。

拉皮德说，以中都是创造了伟大文明的古老民族，现在都致力于加快现代化进程，都拥有强劲的创新能力，双方相互理解、相互欣赏。以方视中国为朋友，建交以来始终奉行一个中国政策，这一立场从未改变，并已成为对华关系的重要基础。以方期待同中方保持密切高层往来，深化各领域合作。

王毅表示，中华民族和犹太民族历史上曾互施援手，今天我们更应相互支持。中方赞赏以方坚持一个中国政策，理解中方维护自身核心利益的正当主张，相信以方将继续支持中方维护主权、安全和发展利益的努力。

王毅说，建交30年来，特别是两国建立创新全面伙伴关系以来，双方各领域合作取得长足发展。中方愿同以方一道，以两国领导人重要共识为指引，充分发挥双方经济互补优势，以创新合作为重要助推器，推动双边关系不断取得更多新成果。

双方一致同意加快推进商签自贸协定，为艰难复苏的全球经济传递正面预期和积极信号。

双方就巴勒斯坦问题交换了意见。王毅表示，中方乐见以方同地区国家建立并发展正常友好关系，这应是中东整体和解的组成部分。以巴和谈长期停滞不前，不符合各方利益，巴勒斯坦同样拥有独立建国的正当权利。中方支持以巴双方以"两国方案"为基础尽快恢复和谈，希望阿拉伯和犹太两大民族能和睦共处，巴以两个国家也能友好相待，实现共同安全。只有共同安全才能带来真正的安全、可持续的安全。中方支持一切有助于实现中东和平的努力，愿为双方直接谈判提供便利。拉皮德表示，以方致力于改善同巴方的关系。

新华网客户端

分享到

基本框架

r.text的内容较多，但真正需要的内容仅占很小一部分，因此有两种方式提取需要的内容：

1) 利用规则

（<p>和 <\p>之间的内容）

<p>和<\p>为段落标记

```
<div class="main-left">
<div id="detail">
<p>    新华社北京4月6日电  4月6日，国务委员兼外长王毅应约同以色列候任总理、外长拉皮德通电话。</p>
<p>    拉皮德说，以中都是创造了伟大文明的古老民族，现在都致力于加快现代化进程，都拥有强劲的创新能力，双方
<p>    王毅表示，中华民族和犹太民族历史上曾互施援手，今天我们更应相互支持。中方赞赏以方坚持一个中国政策，
<p>    王毅说，建交30年来，特别是两国建立创新全面伙伴关系以来，双方各领域合作取得长足发展。中方愿同以方一
<p>    双方一致同意加快推进商签自贸协定，为艰难复苏的全球经济传递正面预期和积极信号。</p>
<p>    双方就巴勒斯坦问题交换了意见。王毅表示，中方乐见以方同地区国家建立并发展正常友好关系，这应是中东整
<p>    双方就乌克兰问题交换意见。王毅介绍了中方的原则立场，表示各方现在都希望实现停火，恢复和平。但要停火
<p>    双方还就伊朗核问题等交换意见。</p>
<div id="articleEdit">
```

基本框架

2) 利用现有的库(BeautifulSoup)

```
import sys
import requests
from bs4 import BeautifulSoup
def get_text(url):
  try:
    news =‘ ’
    r = requests.get(url)
    r.encoding = r.apparent_encoding
```

接下一页PPT

基本框架

2) 利用现有的库(BeautifulSoup)

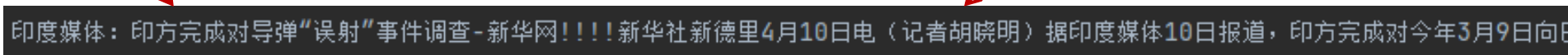
```
    soup = BeautifulSoup(r.text, 'html.parser')
    title = soup.title.text.strip()  ##获得标题
    title +=“!!!!” ##区分标记
    news += title ##
    for x in soup.find_all(‘div’, {‘id’: [‘detail’]}):  ##找到所有div(块),其中id为detail
      for y in x.find_all(‘p’):
        text = y.text.strip() ##得到文本, strip()去除空格
        news += text
    return news
  except
    print("爬取失败")
```

基本框架

2) 利用现有的库(BeautifulSoup)

标题

内容



常见问题

Q1: 如何获取大量数据？

1) 查找url的规律

像百度百科、京东商品等URL是有规律的

```
for i in range(max_num) :
    url = "https://baike.baidu.com/view/" + str(i) + ".htm"
    try:
        text = getHTMLText(url)
    except:
        print(“爬取失败”)
```

常见问题

2）以某个网页(例如新华网国际新闻页面) 为种子，找到该网站所有的超链接（href），再去爬取每个网页

常见问题

2）以某个网页(例如新华网国际新闻页面) 为种子，找到该网站所有的超链接（href），再去爬取每个网页

```
from bs4 import BeautifulSoup  ##导入BeautifulSoup库

def get_all_url(url):
    try:
        news_list = []        ##空列表
        r = requests.get(url) ## 解析种子网页
        r.encoding = r.apparent_encoding
        soup = BeautifulSoup(r.text, ‘html.parser’) ##利用BeautifulSoup库解析
        tags = soup.find_all( ‘a’ ) ##找到所有锚/超链接
        for tag in tags:
            news_list.append((str(tag.get( ‘href’ )).strip())) ##得到href
        return news_list
    except:
        print("爬取失败")
```

常见问题

2）以某个网页(例如新华网国际新闻页面) 为种子，找到该网站所有的超链接（**href**），再去爬取每个网页

news_list=['http://www.news.cn/2022-04/11/c_1128550818.htm',
'http://www.news.cn/2022-04/11/c_1128550818.htm',
'http://www.news.cn/2022-04/11/c_1128550726.htm',
'http://www.news.cn/2022-04/11/c_1211635467.htm',
'http://www.news.cn/2022-04/11/c_1128550760.htm',
'http://www.news.cn/world/2022-04/11/c_1128548373.htm',
'http://www.news.cn/world/2022-04/10/c_1211635017.htm',
'http://www.news.cn/world/2022-04/08/c_1211634245.htm',
'http://www.news.cn/world/2022-04/08/c_1211634246.htm’ ,
…]

常见问题

整体程序

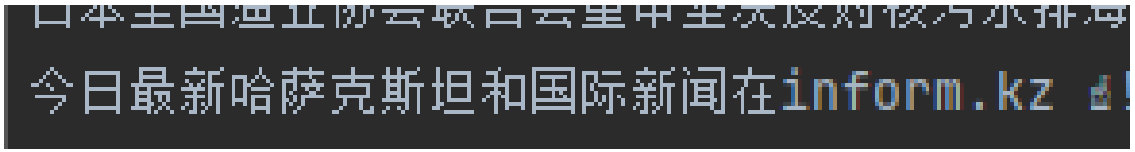
```
def get_xinhua_news(home_url):
    path=‘xinhua_news.txt’  ##文件名字
    news_list = [] ##list
    url_list=get_all_url(home_url)##得到所有url列表
    url_list=list(set(url_list))##去重
    for url in url_list:
        news=get_text(url)##得到每个url的内容
        if news==None: ##去掉为空的url
          continue
        news_list.append(news)##添加到同一个list中
        write_txt(news_list, path)##写文件

if __name__ == '__main__':
    home_url = "http://www.xinhuanet.com/worldpro/"
    get_xinhua_news(home_url)
```

常见问题

Q2: 为何要做数据清洗？

爬取的数据里面会有一部分噪声



Thanks