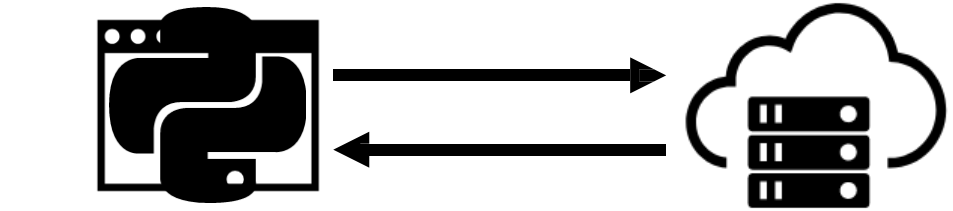
Knowledge Discovery and Data Mining

Lab 2 Introduction to Python Data Crawler

Tianyue Zheng
zhengty@sustech.edu.cn

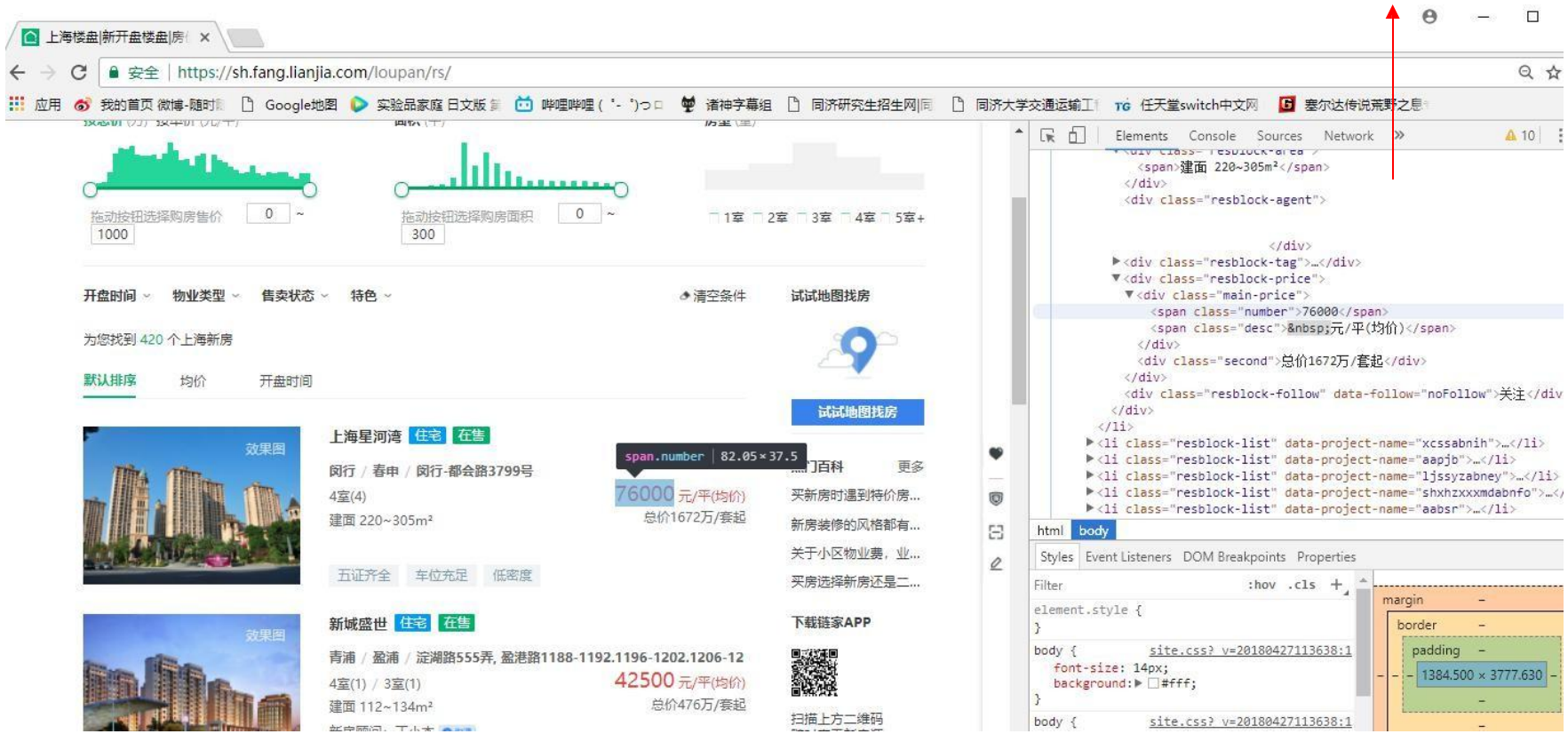
数据获取方式：爬虫



Retrieved from: 余庆，李玮峰《交通时空大数据分析、挖掘与可视化》

爬虫1.0： html抓取

网页源代码html



写个Python循环，获取每条记录的房价

Retrieved from: 余庆，李玮峰《交通时空大数据分析、挖掘与可视化》

爬虫2.0：API接口

百度地图开放平台

高德开放平台 lbs.amap.com

新浪微博开放平台

注册为开发者

获取key

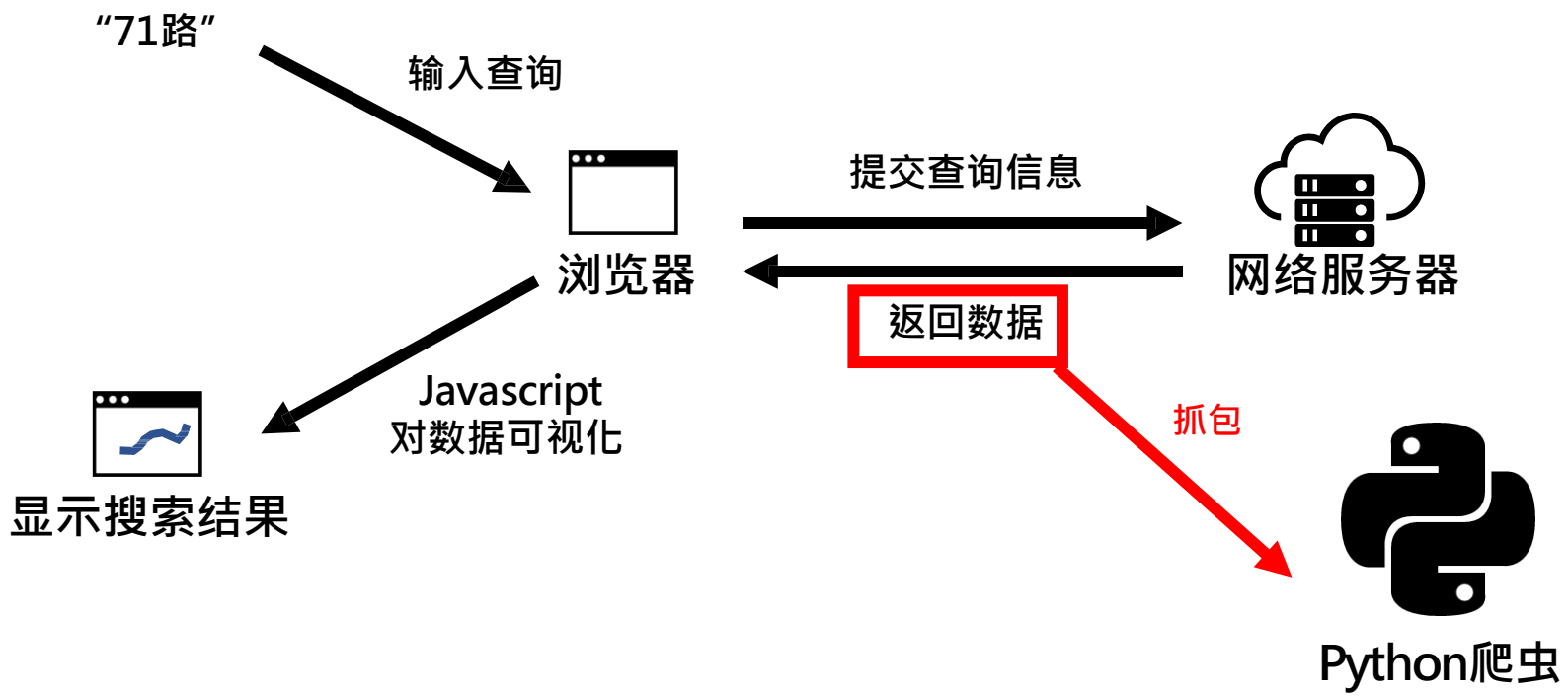
向服务器提交查询
（附带key）

获得数据

Retrieved from: 余庆，李玮峰《交通时空大数据分析、挖掘与可视化》

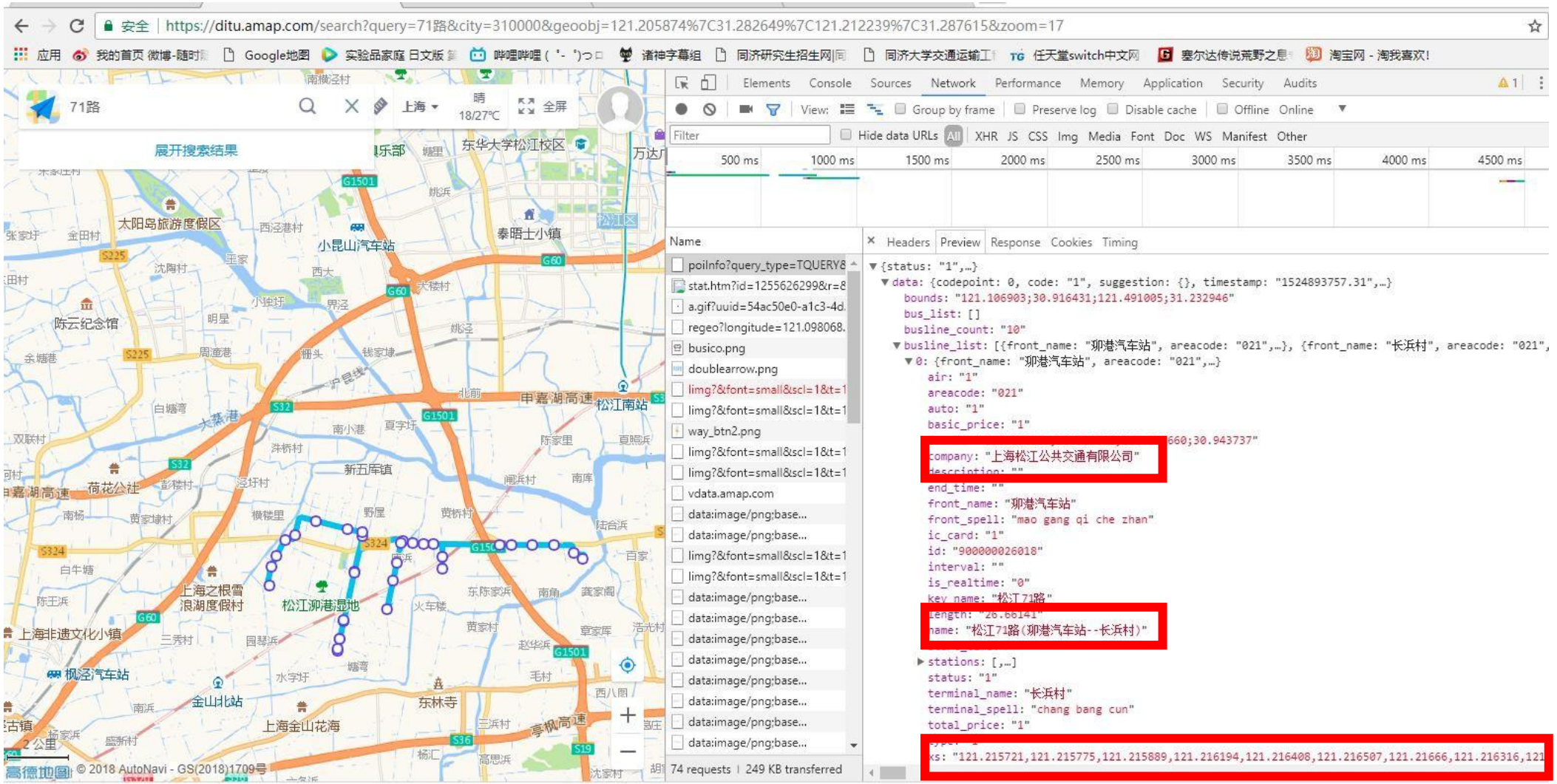
爬虫3.0：抓包

在百度地图上搜索公交线“71路”：



Retrieved from: 余庆，李玮峰《交通时空大数据分析、挖掘与可视化》

爬虫3.0：抓包


https://ditu.amap.com/search?query=71路&city=310000&geoobj=121.205874%7C31.282649%7C121.212239%7C31.287615&zoom=17
71路
展开搜索结果
Elements Console Sources Network Performance Memory Application Security Audits
Name
Headers Preview Response Cookies Timing
"company": "上海松江公共交通有限公司"
"name": "松江71路(聊港汽车站--长浜村)"
74 requests | 249 KB transferred

爬虫能够获取到什么数据?

兴趣点数据

房价数据

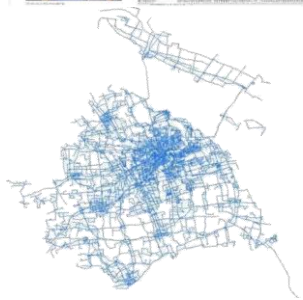
路径规划

等时圈数据

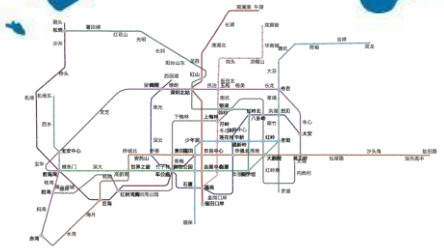
微博数据

行政区划

公交线路



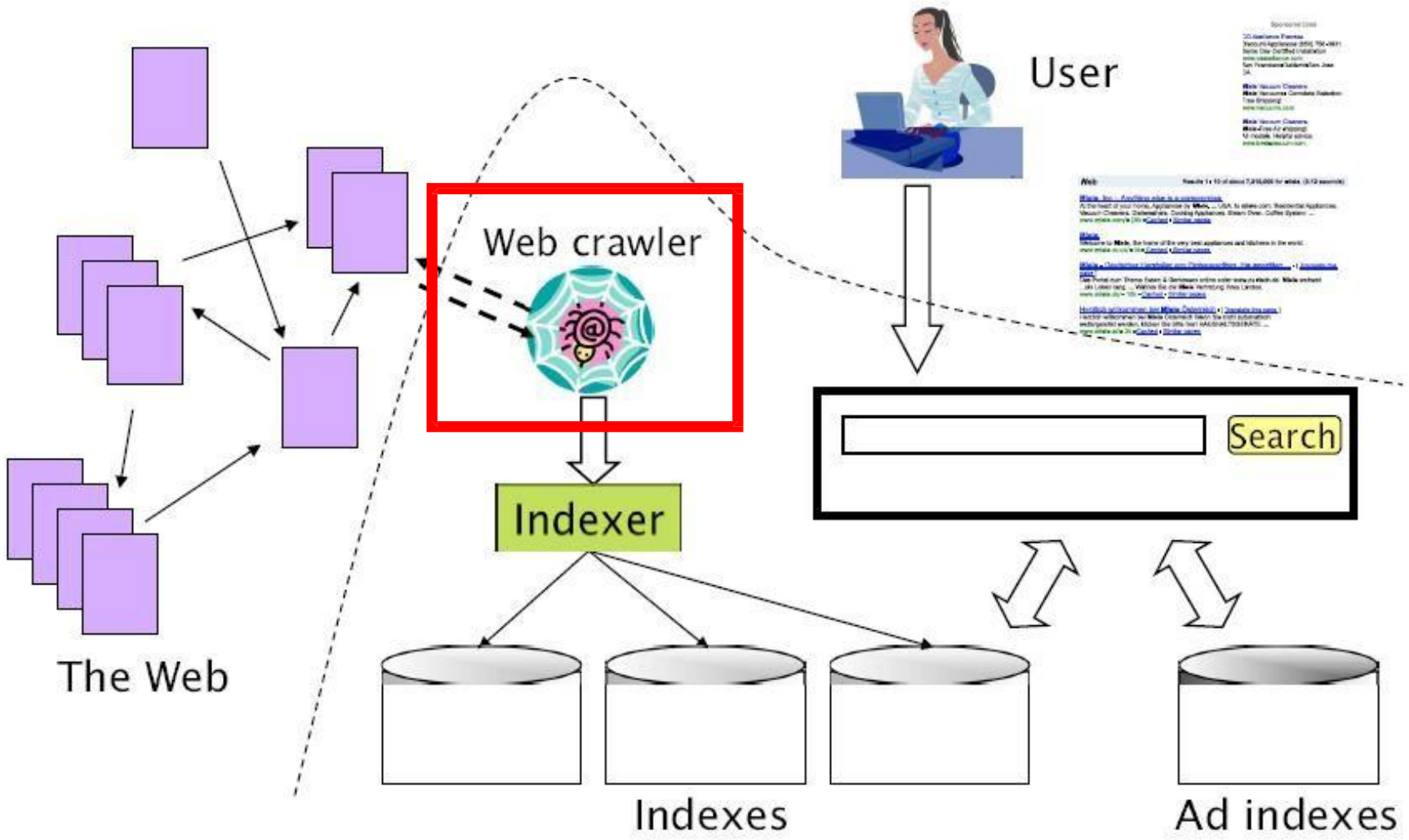
地铁线路



Retrieved from: 余庆，李玮峰《交通时空大数据分析、挖掘与可视化》

Web Crawler

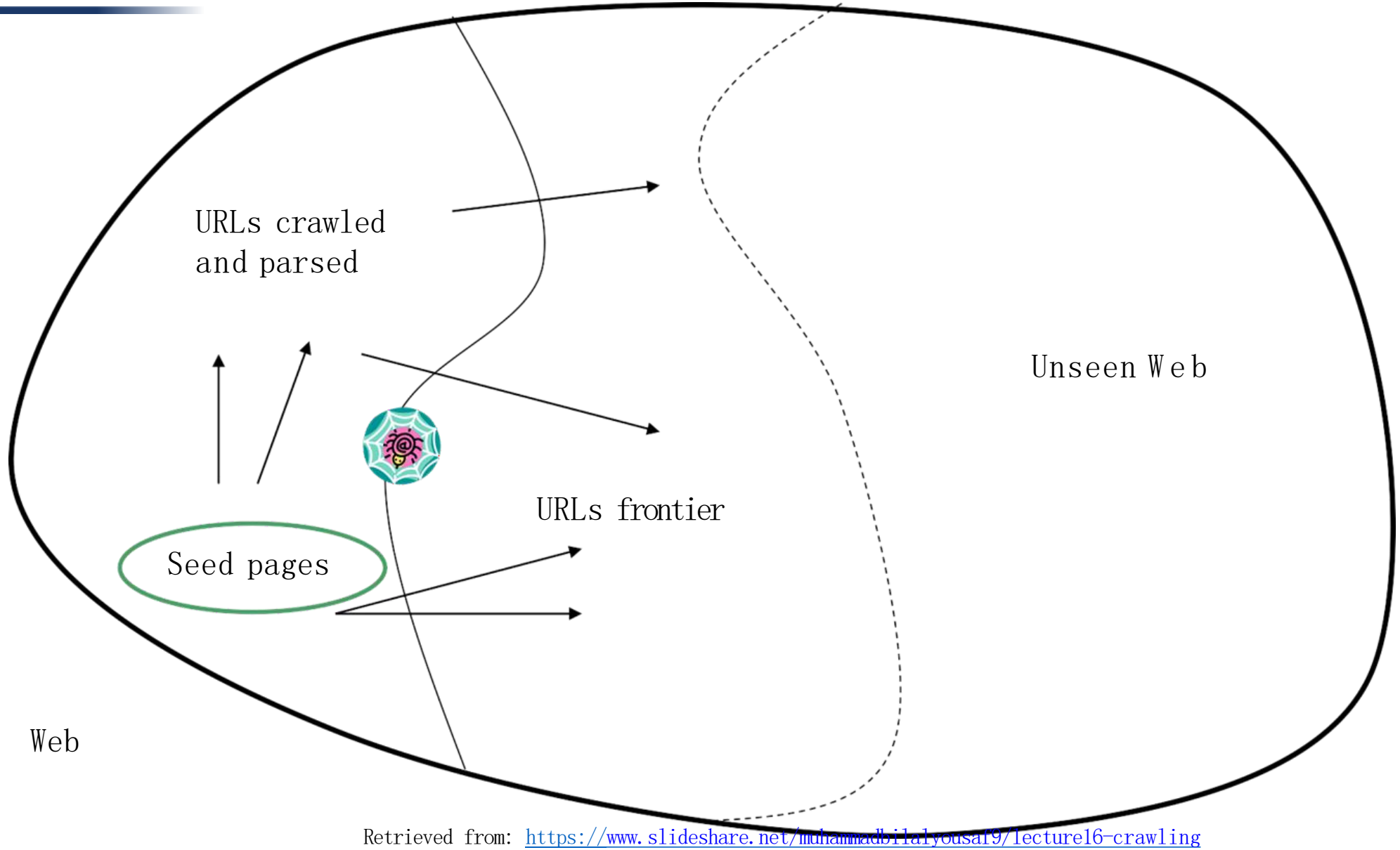
The Crawler



▶ Figure 19.7 The various components of a web search engine.

How hard can crawling be?

Basic Crawler Operations



Retrieved from: https://www.slideshare.net/muhammadbilalyousaf9/lecture16-crawling

Simple Crawler

```
urlqueue := (some carefully selected set of seed urls)

while urlqueue is not empty:
   myurl := urlqueue.getlastanddelete()
   mypage := myurl.fetch()
   fetchedurls.add(myurl)
   newurls := mypage.extracturls()

   for myurl in newurls:
      if myurl not in fetchedurls and not in urlqueue:
         urlqueue.add(myurl)

   indexer.index(mypage)
```

What's wrong with this crawler?

Features a crawler MUST provide

Robustness

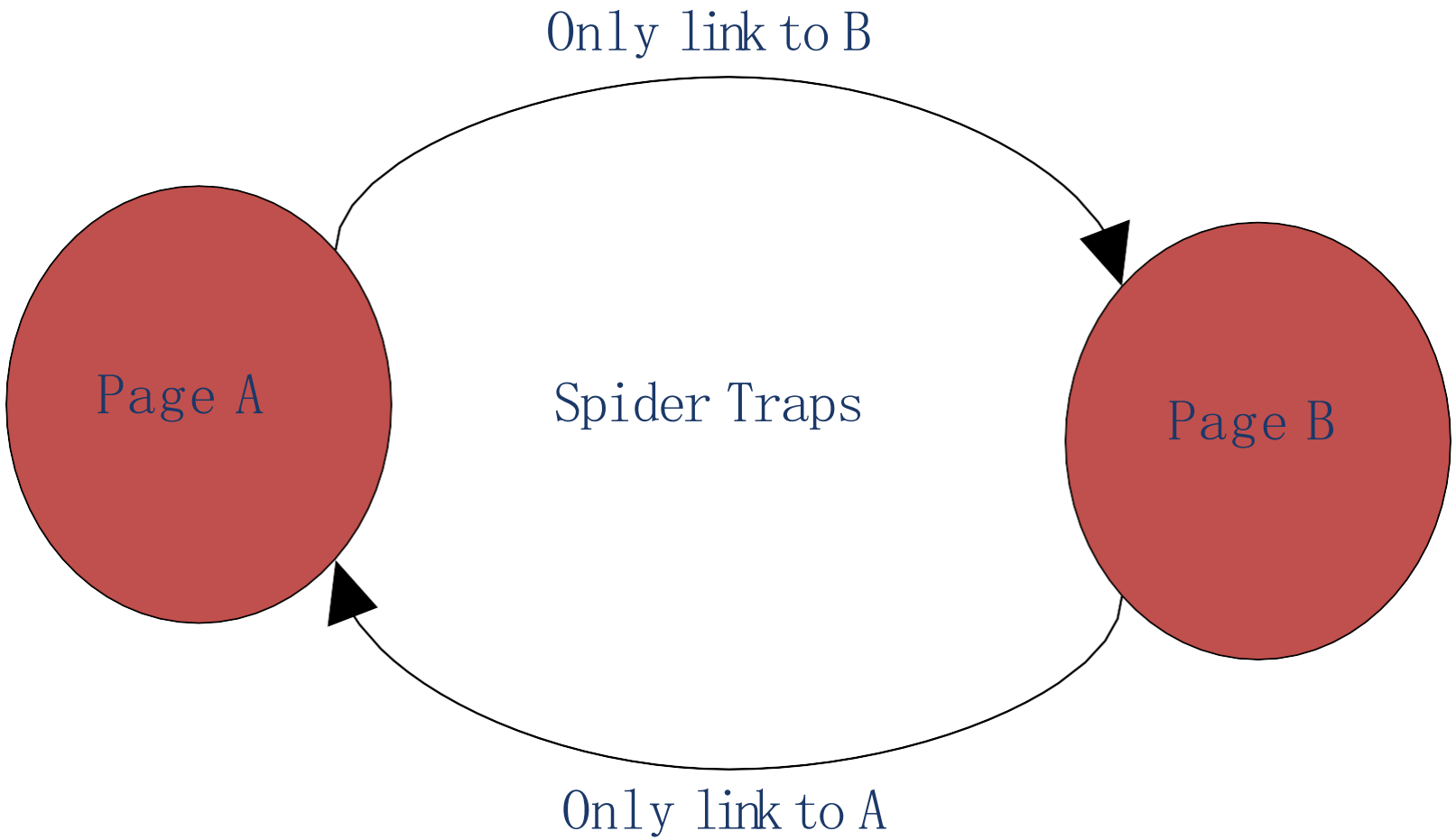
Politeness

Politeness

1) Robots.txt

2) Do not frequently request the same site

Robustness



Features a crawler SHOULD provide

Distributed

Scalable

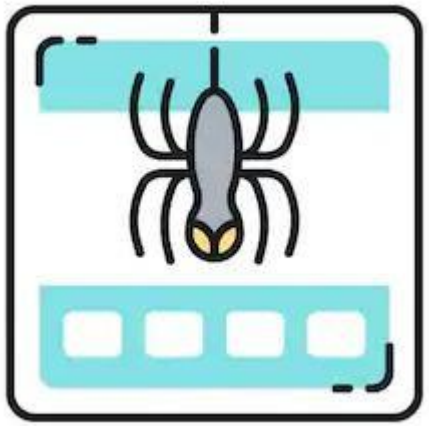
Performance and efficiency

Quality

Freshness

Extensible

What we do today



Quotes to Scrape

Login

"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking."
by **Albert Einstein** (about)
Tags: change deep-thoughts thinking world

"It is our choices, Harry, that show what we truly are, far more than our abilities."
by **J.K. Rowling** (about)
Tags: abilities choices

"There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle."
by **Albert Einstein** (about)
Tags: inspirational life live miracle miracles

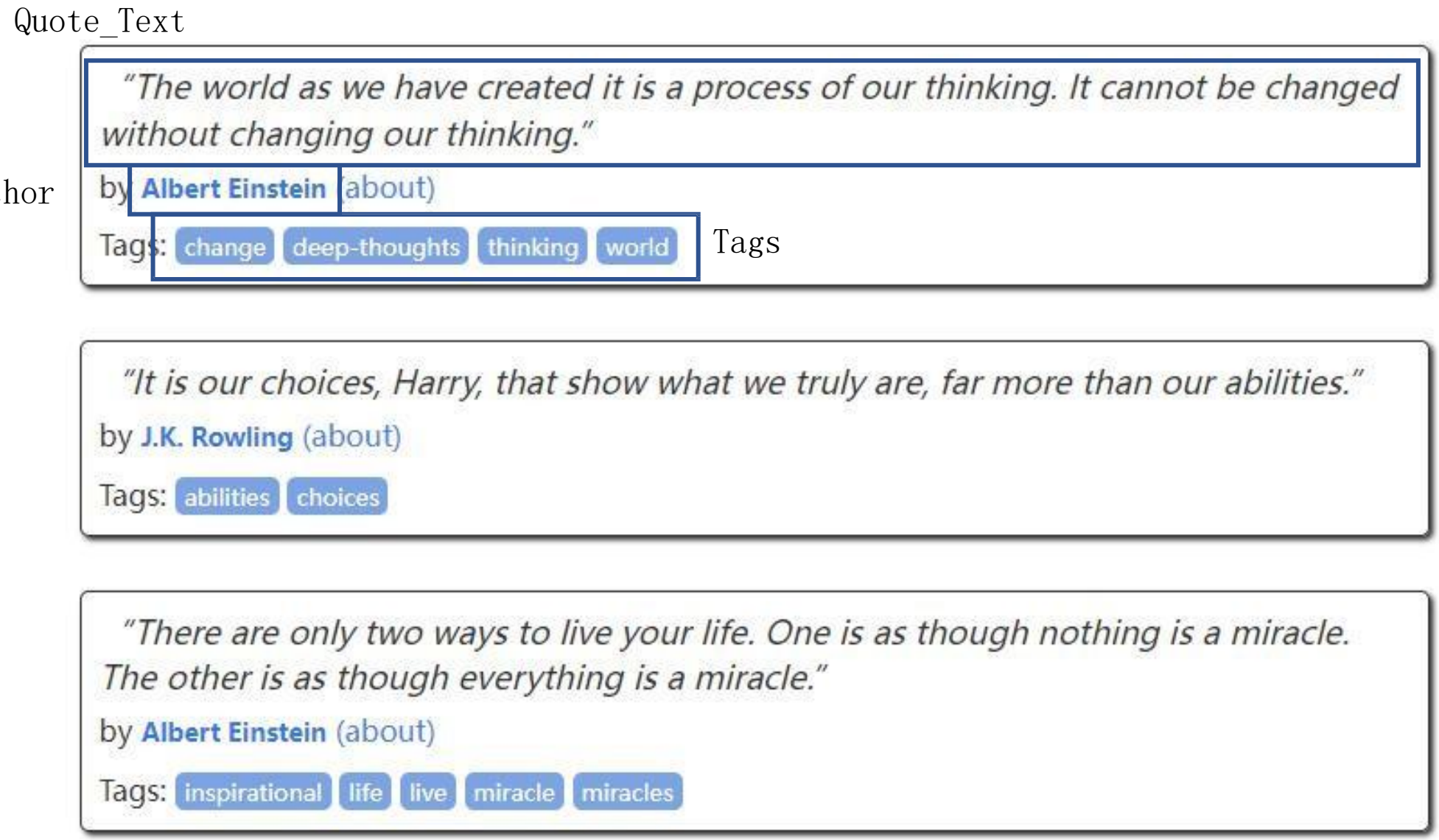
"The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid."
by **Jane Austen** (about)
Tags: aliteracy books classic humor

"Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring."
by **Marilyn Monroe** (about)
Tags: be-yourself inspirational

Top Ten tags

love
inspirational
life
humor
books
reading
friendship
friends
truth
simile

What we want today



Tools

Beautiful Soup 4

https://www.crummy.com/software/BeautifulSoup/bs4/doc.zh/#

Setup

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
url = "https://quotes.toscrape.com/"
html = urlopen(url)
soup = BeautifulSoup(html, 'html.parser')
type(soup)
# Print out the text
text = soup.get_text()
#print(soup.text)
```

Exercise

Get the quote_text, author and tags from every post on the Quotes to Scrape site.


```
[{'author': 'Albert Einstein', 'quote_text': '“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”', 'tags': ['change', 'deep-thoughts', 't
hinking', 'world']}, {'author': 'J.K. Rowling', 'quote_text': '“It is our choices, Harry, that show what we truly are, far more than our abilities.”', 'tags': ['abilities', 'choices']}, {'author': 'Albe
rt Einstein', 'quote_text': '“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”', 'tags': ['inspirational', 'life', 'liv
e', 'miracle', 'miracles']}, {'author': 'Jane Austen', 'quote_text': '“The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.”', 'tags': ['aliteracy', 'bo
oks', 'classic', 'humor']}, {'author': 'Marilyn Monroe', 'quote_text': "“Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring.”", 'tags': ['be-you
rself', 'inspirational']}, {'author': 'Albert Einstein', 'quote_text': '“Try not to become a man of success. Rather become a man of value.”', 'tags': ['adulthood', 'success', 'value']}, {'author': 'Andr
é Gide', 'quote_text': '“It is better to be hated for what you are than to be loved for what you are not.”', 'tags': ['life', 'love']}, {'author': 'Thomas A. Edison', 'quote_text': "“I have not failed.
I've just found 10,000 ways that won't work.”", 'tags': ['edison', 'failure', 'inspirational', 'paraphrased']}, {'author': 'Eleanor Roosevelt', 'quote_text': "“A woman is like a tea bag; you never know
how strong it is until it's in hot water.”", 'tags': ['misattributed-eleanor-roosevelt']}, {'author': 'Steve Martin', 'quote_text': '“A day without sunshine is like, you know, night.”', 'tags': ['humo
r', 'obvious', 'simile']}, {'author': 'Marilyn Monroe', 'quote_text': "“This life is what you make it. No matter what, you're going to mess up sometimes, it's a universal truth. But the good part is you
get to decide how you're going to mess it up. Girls will be your friends - they'll act like it anyway. But just remember, some come, some go. The ones that stay with you through everything - they're you
r true best friends. Don't let go of them. Also remember, sisters make the best friends in the world. As for lovers, well, they'll come and go too. And baby, I hate to say it, most of them - actually pr
```

Hint

- Hint 1: Understand the html structure of the page can be very helpful!
- Hint 2: You can use BeautifulSoup to find css element to pin- point what you need.
- Hint 3: You can also grab the url for next page to recursively scrape the whole site.

End of Lab 2