

# Lesson 6

Crawler

# Ex

1. Nhập n, in ra từ 0 đến n-1:

```
Enter a number: 17
```

```
0 1 2 3 4 5 6 7 8 9 10 11
```

2. Nhập n, in ra bảng sau

```
Enter a number: 10
```

1	2	3	4	5	6	7	8	9	10
2	4	6	8	10	12	14	16	18	20
3	6	9	12	15	18	21	24	27	30
4	8	12	16	20	24	28	32	36	40
5	10	15	20	25	30	35	40	45	50
6	12	18	24	30	36	42	48	54	60
7	14	21	28	35	42	49	56	63	70
8	16	24	32	40	48	56	64	72	80
9	18	27	36	45	54	63	72	81	90
10	20	30	40	50	60	70	80	90	100

# Data

<http://s.cafef.vn/bao-cai-tai-chinh/VNM/IncSta/2018/0/0/0/ket-qua-hoat-dong-kinh-doanh-cong-ty-co-phan-sua-viet-nam.chn>

<https://www.niche.com/>

<https://www.usnews.com/best-colleges>

<https://dantri.com.vn/>

<https://www.youtube.com>

[https://vnreview.vn/tin-tuc-xa-hoi-so/-/view\\_content/content/2497592/chung-ta-tao-ra-bao-nhieu-du-lieu-moi-ngay](https://vnreview.vn/tin-tuc-xa-hoi-so/-/view_content/content/2497592/chung-ta-tao-ra-bao-nhieu-du-lieu-moi-ngay)

# Data

- Dữ liệu trong xã hội loài người đã được số hóa phần lớn
- Internet chứa gần như mọi loại dữ liệu thông thường: Video, ảnh, thông tin nhà đất, thời tiết, y học,...
- Một số thông tin là tĩnh, một số thông tin là động, một số realtime
- Thông tin quá nhiều nhưng cho đến nay vẫn chưa có công cụ nào khai thác được toàn bộ các dữ liệu này

Khó khăn:

- Dữ liệu quá lớn
- Các trung tâm dữ liệu không chia sẻ
- Đa số các thông tin ở dạng phi cấu trúc, khó dùng được ngay
- ...

# Crawler - Là gì?

- Là một công cụ (phần mềm, modules, plugins hoặc đơn giản là 1 function)
- Tự động bóc tách, phân tích dữ liệu từ nguồn

# Crawler - Dùng làm gì?

**Hacker News** | new | threads | comments | show | ask | jobs | submit

1. **'Extreme poverty' to fall below 10% of world population for first time** (theguardian.com)  
68 points by hlyan 2 hours ago | flag | 68 comments | save to pocket | add to buffer
2. **Inside the Creation of the Microsoft Surface Book** (mashable.com)  
48 points by \_nh\_ 1 hour ago | flag | 50 comments | save to pocket | add to buffer
3. **FBI director calls lack of data on police shootings "ridiculous," "embarrassing"** (washingtonpost.com)  
15 points by jwvns 27 minutes ago | flag | 3 comments | save to pocket | add to buffer
4. **Startup NASA** (nasa.gov)  
30 points by rusesaerensen 1 hour ago | flag | 4 comments | save to pocket | add to buffer
5. **Doubling the speed of jpegtran with SIMD** (cloudflare.com)  
79 points by grahamc 3 hours ago | flag | 14 comments | save to pocket | add to buffer
6. **Winklevoss Twins' Bitcoin Exchange** (qemini.com)  
78 points by wehadfun 2 hours ago | flag | 34 comments | save to pocket | add to buffer
7. **OpenPGP SEIP downgrade attack** (metzdowd.com)  
23 points by mulyu 1 hour ago | flag | 2 comments | save to pocket | add to buffer
8. **How a small streaming site became the Netflix for indie film** (theverge.com)  
24 points by entor 1 hour ago | flag | 6 comments | save to pocket | add to buffer
9. **The 1810 Republic of West Florida** (vox.com)  
7 points by devicberker 34 minutes ago | flag | 2 comments | save to pocket | add to buffer
10. **Thinking in GraphQL** (facebook.github.io)  
54 points by bshkr 2 hours ago | flag | 20 comments | save to pocket | add to buffer
11. **Seattle, in Midst of Tech Boom, Tries to Keep Its Soul** (nytimes.com)  
4 points by vanderhage 28 minutes ago | flag | discuss | save to pocket | add to buffer
12. **The Nobel Prize in Literature 2015** (nobelprize.org)  
26 points by danielch 2 hours ago | flag | 6 comments | save to pocket | add to buffer



```
{
  "content": [
    {
      "title": {
        "text": "'Extreme poverty'
to fall below 10% of world
population for first time",
        "href": "http://
www.theguardian.com/society/2015/
oct/05/world-bank-extreme-poverty-
to-fall-below-10-of-world-
population-for-first-time"
      },
      "points": "9 points",
    }
  ]
}
```

# Crawler - Dùng làm gì?

- Sử dụng để tạo ra các trang web tổng hợp dữ liệu từ các nguồn khác (so sánh giá, tăng cạnh tranh, phân tích đối thủ):

<https://magiamgia.guru/lich-su-gia-ban/>

- Sử dụng để lấy dữ liệu, phục vụ cho các mục đích phân tích chuyên sâu

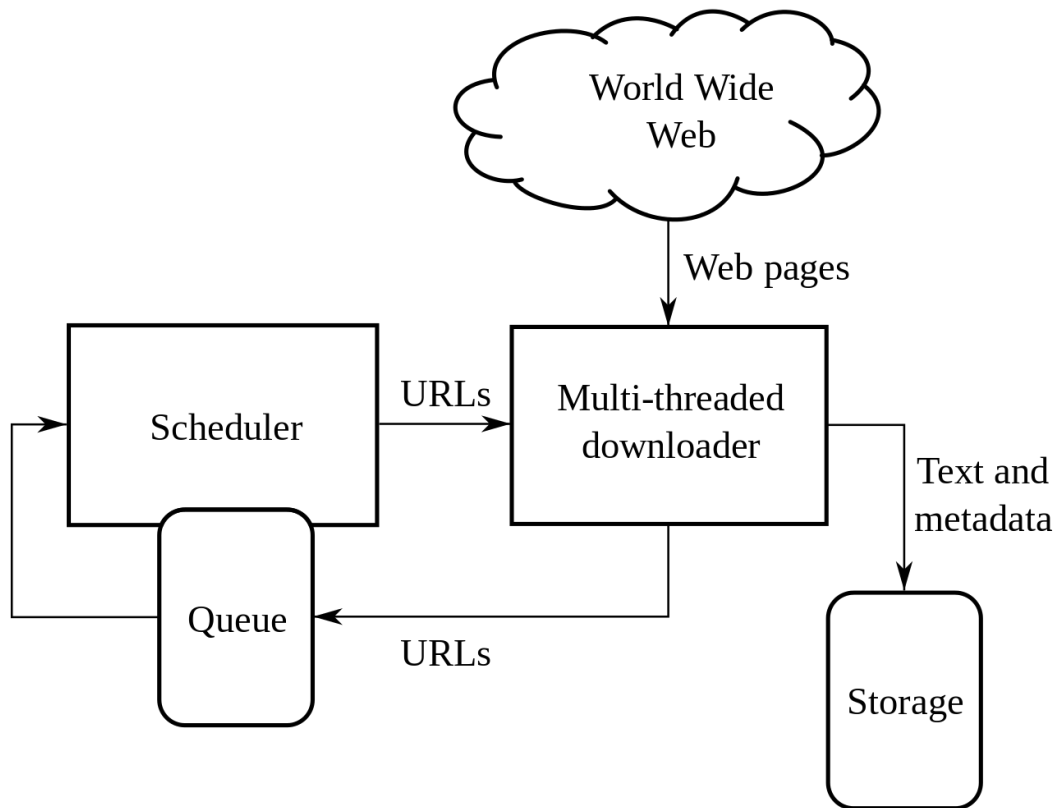
<http://viet.jnlp.org/download-du-lieu-tu-vung-corpus>

- Sử dụng trong lĩnh vực seo để tự động lấy thông tin từ các site khác

<https://kdnautoleech.com>

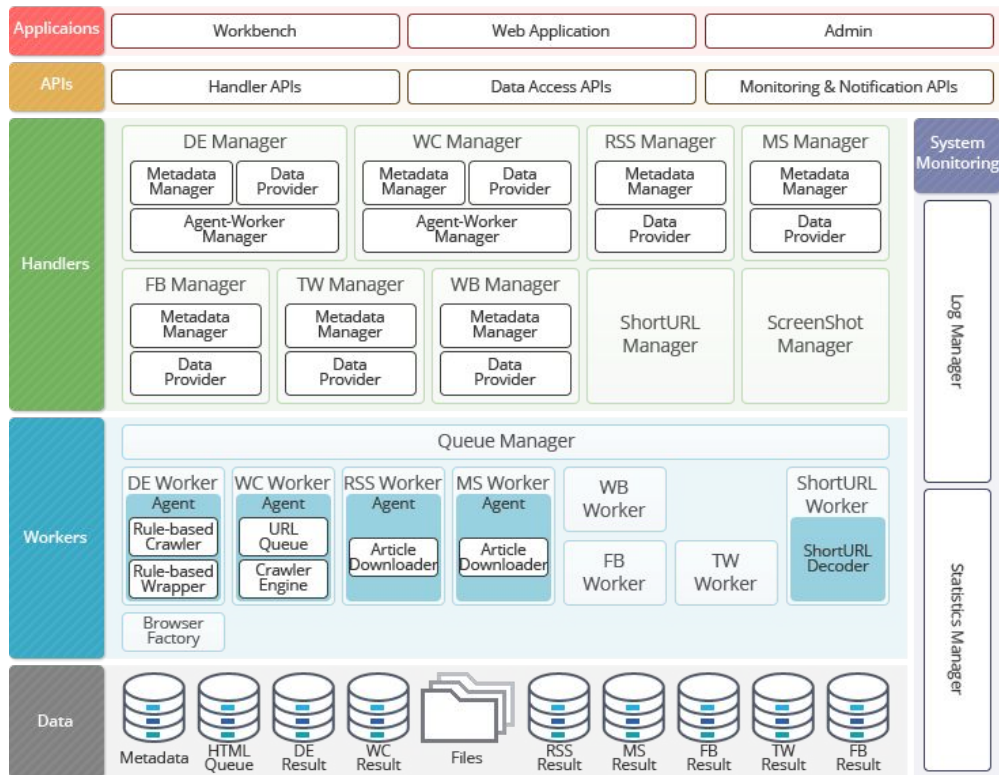
- ...project cuối khóa.

# Crawler - Hoạt động như thế nào?





# Crawler - Hoạt động như thế nào?



# Crawler - Lấy HTML

**Mục tiêu:** Crawl mục sức khỏe của dantri.vn

**Các bước thực hiện:**

- Lấy html
- Bóc tách dữ liệu
- Lưu trữ dữ liệu

# Crawler - Làm thế nào?

- HTML là gì?
- Lấy HTML từ trang web cần lấy
- + chrome: view source
- + python: requests

# Crawler - Làm thế nào?

## - HTML là gì?

- + HyperText Markup Language, hay là "Ngôn ngữ Đánh dấu Siêu văn bản".
- + Được thiết kế ra để tạo nên các trang web với các mẫu thông tin được trình bày trên World Wide Web.
- + WWW - mạng lưới toàn cầu là một không gian thông tin toàn cầu mà mọi người có thể truy cập (đọc và viết) thông tin qua các thiết bị kết nối với mạng Internet.
- + HTML không phải là ngôn ngữ lập trình.

# Crawler - Làm thế nào?

```
<!DOCTYPE html>
<html>
<body>

<h1>My First Heading</h1>

<p>My first paragraph.</p>

</body>
</html>
```

## My First Heading

My first paragraph.

**xin chào**

# Crawler - Làm thế nào?

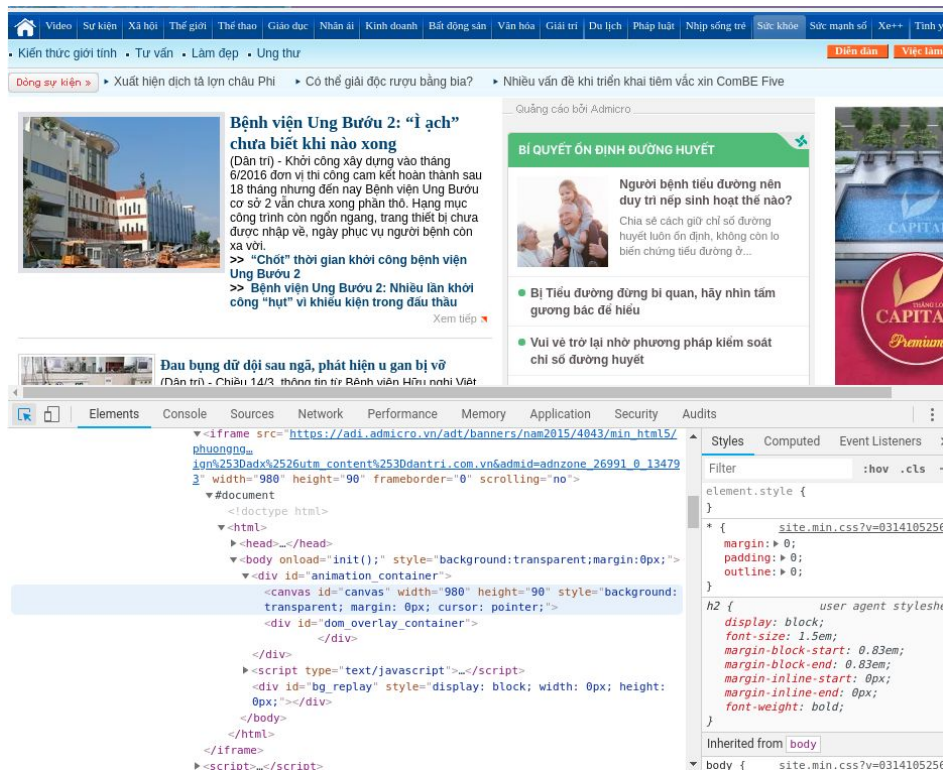
Tìm hiểu về một số thẻ HTML ? (Tạo file index.html để thử nghiệm)

- Thẻ html
- Thẻ head, title
- Thẻ body, p
- Thẻ a - thuộc tính href
- Thẻ img - thuộc tính src

# Crawler - Làm thế nào?

+ Chrome view source

+ Chrome developer



# Crawler - Làm thế nào?

Lấy html bằng python

**import** requests

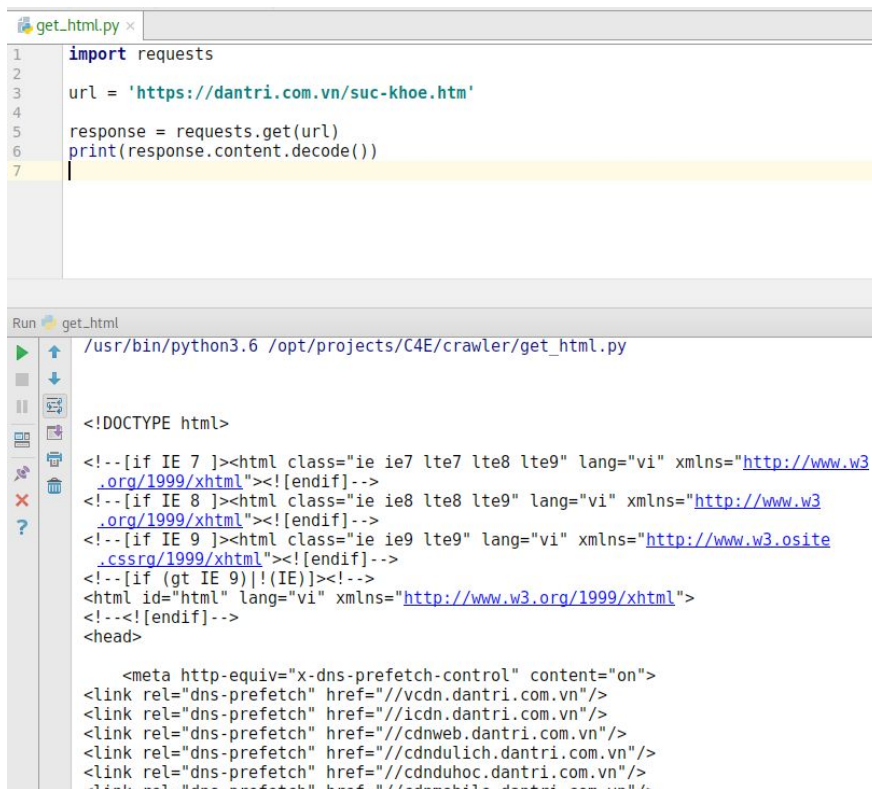
url = **'https://dantri.com.vn/suc-khoe.htm'**

response = requests.get(url)

**print**(response.content.decode())

**with** open('a.html', 'wt') **as** f:

f.write(response.content.decode())



The screenshot shows a code editor with a file named `get_html.py` containing the following Python code:

```
1 import requests
2
3 url = 'https://dantri.com.vn/suc-khoe.htm'
4
5 response = requests.get(url)
6 print(response.content.decode())
7
```

Below the code editor, a terminal window titled `Run get_html` displays the output of the script, which is the HTML content of the specified URL:

```
/usr/bin/python3.6 /opt/projects/C4E/crawler/get_html.py

<!DOCTYPE html>

<!--[if IE 7 ]><html class="ie ie7 lte7 lte8 lte9" lang="vi" xmlns="http://www.w3
.org/1999/xhtml"><![endif]-->
<!--[if IE 8 ]><html class="ie ie8 lte8 lte9" lang="vi" xmlns="http://www.w3
.org/1999/xhtml"><![endif]-->
<!--[if IE 9 ]><html class="ie ie9 lte9" lang="vi" xmlns="http://www.w3.org
.cssrg/1999/xhtml"><![endif]-->
<!--[if (gt IE 9)|!(IE)]><!-->
<html id="html" lang="vi" xmlns="http://www.w3.org/1999/xhtml">
<!--<![endif]-->
<head>

    <meta http-equiv="x-dns-prefetch-control" content="on">
<link rel="dns-prefetch" href="//vcdn.dantri.com.vn"/>
<link rel="dns-prefetch" href="//icdn.dantri.com.vn"/>
<link rel="dns-prefetch" href="//cdnweb.dantri.com.vn"/>
<link rel="dns-prefetch" href="//cdndulich.dantri.com.vn"/>
<link rel="dns-prefetch" href="//cdnduhoc.dantri.com.vn"/>
<link rel="dns-prefetch" href="//cdnmobile.dantri.com.vn"/>
```



# Crawler - Làm thế nào?

Dữ liệu html đã lấy chứa tất các thông tin trang web hiển thị, bao gồm cả những không cần lấy như: menu, quảng cáo... => Cần bóc tách đúng phần danh sách tin tức



## **Đau bụng dữ dội sau ngã, phát hiện u gan bị vỡ**

(Dân trí) - Chiều 14/3, thông tin từ Bệnh viện Hữu nghị Việt Nam – Cu Ba Đồng Hới cho biết, các bác sĩ tại bệnh viện này vừa cấp cứu thành công cho một bệnh nhân bị vỡ u gan, hơn 2 lít máu tràn ổ bụng.

[Xem tiếp](#)



## **Nguyên nhân dịch tả lợn châu Phi lây lan ra các địa phương**

(Dân trí) - Cục Thú y cho biết, một số người chưa nhận thức được tính chất nguy hiểm của dịch tả lợn châu Phi, cũng như vì lợi ích kinh tế trước mắt nên khi có lợn bệnh, lợn chết đã mua bán, vận chuyển, giết mổ, tiêu thụ lợn bệnh, lợn nghi mắc bệnh, dẫn đến dịch bệnh lây lan nhanh, ở phạm vi rộng.  
>> **Dịch tả lợn châu Phi đã lan ra 17 tỉnh, thành phố của Việt Nam**

[Xem tiếp](#)



## **Vinmec triển khai liệu pháp miễn dịch tự thân và nhiệt trị kết hợp điều trị ung thư**

(Dân trí) - Vinmec vừa trở thành bệnh viện đầu tiên tại Việt Nam triển khai liệu pháp điều trị ung thư hiện đại là miễn dịch tự thân và nhiệt trị. Đây là phương pháp mới, có hiệu quả điều trị cao, chống tái phát tốt, đang được áp dụng thành công tại các quốc gia có nền y học phát triển như Nhật Bản, Mỹ và châu Âu

[Xem tiếp](#)

# Crawler - I

Beautiful soup: l  
xuất dữ liệu từ c

- Cài đặt Bea

pip3 install l

- Tìm vùng q
- Trích rút th

<Sử dụng pytho

```
2
3 html = """
4 <!DOCTYPE html>
5 <html>
6 <title>Trang web của tôi</title>
7
8 <body>
9     <h1>Chào mừng đến với Techkids</h1>
10     <p id='data'>Xin chào</p>
11     <p>python</p>
12     <p>c#</p>
13     <p>java</p>
14 </body>
15
16 </html>
17 """
18
19 soup = BeautifulSoup(html, 'html.parser')
20 print("title:", soup.title)
21 print("title string:", soup.title.string)
22
23 print("p attribute:", soup.p.attrs)
24 print("p id:", soup.p['id'])
```

# Crawler - Làm thế nào?

- Áp dụng để lấy danh sách các bài vi

```
crawl_dantri.py x
1 import requests
2 from bs4 import BeautifulSoup
3
4 url = 'https://dantri.com.vn/suc-khoe.htm'
5
6 response = requests.get(url)
7 soup = BeautifulSoup(response.content, 'html.parser')
8
9 # Cách 1. Tìm các bài viết là các con của thẻ listcheckepl
10 # lấy element chứa các bài viết
11 post_elements = soup.find(id='listcheckepl')
12 # Lấy danh sách các bài viết con
13 all_post = list(post_elements.children)
14 # print(all_post) # => Ra nhiều các thẻ rác, sử dụng cách bên dưới
15
16 # cách 2. Tìm bài viết theo điều kiện
17 post_elements = soup.find_all("div", {"data-boxtype": 'timelineposition'})
18 for v in post_elements:
19     print(v.a.attrs['title'])
20     print(v.a.img.attrs['src'])
21
```

# Crawler - Làm thế nào?

Bóc tách danh sách các bài viết

Cách 1: Tìm các bài viết là các con của thẻ listckekepl

```
post_elments = soup.find(id='listckekepl')
all_post = list(post_elments.children)
```

Cách 2: Tìm bài viết theo điều kiện

```
18 post_elments = soup.find_all("div", {"data-boxtype": 'timelineposition'})
19
20 for v in post_elments:
21     print(v.a.attrs['title'])
22     print(v.a.img.attrs['src'])
```

# Crawler - Làm thế nào?

Lưu trữ dữ liệu vào file json

```
import json
```

```
# result là dữ liệu đã trích rút được
```

```
with open('result.json', 'wt', encoding='utf-8') as f:
```

```
    f.write(json.dumps(result, ensure_ascii=False))
```

# Crawler - Làm thế nào?

Lưu trữ dữ liệu vào excel: `python -m pip install pyexcel pyexcel-xls`

```
1 import pyexcel as p
2 arr = [{'title': 'title1', 'content': 'content1'},
3        {'title': 'title2', 'content': 'content2'}]
4
5 p.save_as(records=arr, dest_file_name="result.xls")
6
```

```
25 arr=[]
26 for v in post_elments:
27     arr.append([v.a.attrs['title'],v.a.img.attrs['src']])
28
29 p.save_as(array=arr,dest_file_name= 'dantri.xls')
```

# Hướng phát triển

- Lấy từ nhiều nguồn khác nhau => Cấu trúc html khác nhau => ?
- Chạy phân tán trên nhiều máy
- Xử lý dữ liệu thu được: Post lên 1 website, Mining trên dữ liệu
- Xử lý các trang phức tạp