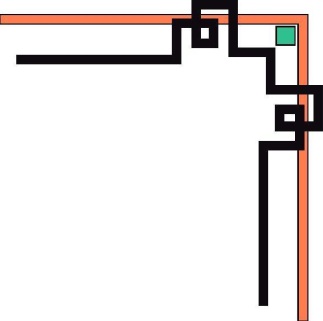
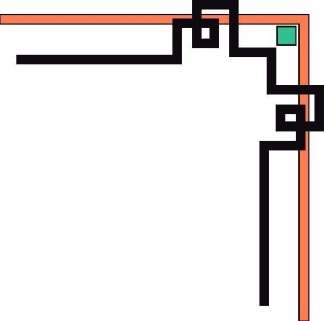
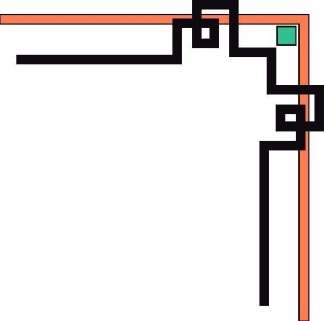
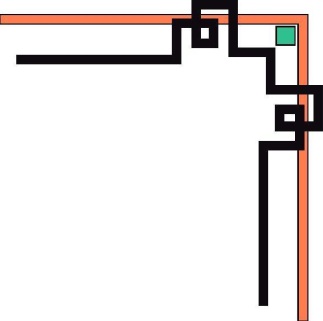
**BỘ GIÁO DỤC VÀ ĐÀO TẠO**



**TRƯỜNG ĐẠI HỌC KIẾN TRÚC ĐÀ NẲNG**

**KHOA CÔNG NGHỆ**



**ĐỀ TÀI**

**TÌM HIỂU ỨNG DỤNG CRAWL**

**LẤY DỮ LIỆU VÀ TIN TỨC TỪ CÁC WEBSITE**



**NHÓM : Nguyễn Văn Rô**

**Trần Minh Tiến**

**LỚP : 16 CT**

**GVHD : Đỗ Phúc Hảo**

* **Đà Nẵng, tháng 05 năm 2019**  -

Mục lục

[CHƯƠNG I : GIỚI THIỆU 3](#_Toc9981482)

[I. Mở đầu 3](#_Toc9981483)

[1. Yêu cầu và mục tiêu của đề tài 3](#_Toc9981484)

[1.1 Yêu cầu 3](#_Toc9981485)

[1.2 Mục tiêu 3](#_Toc9981486)

[CHƯƠNG II. TÌM HIỂU VÀ PHÂN TÍCH SƠ BỘ 4](#_Toc9981487)

[I . Tìm hiểu các kiến thức tổng quan 4](#_Toc9981488)

[1. World Wide Web 4](#_Toc9981489)

[2. Web Crawler 4](#_Toc9981490)

[3. Web Scraper 5](#_Toc9981491)

[II. Phân loại Web 5](#_Toc9981492)

[1. Phân loại dựa vào sự thể hiện của tính động 5](#_Toc9981493)

[2. Phân loại dựa vào cách thức tạo ra tính động 6](#_Toc9981494)

[CHƯƠNG III. CÁC KIẾN THỨC NỀN TẢNG 6](#_Toc9981495)

[I . HTML 6](#_Toc9981496)

[II . XML 6](#_Toc9981497)

[III . XHTML 7](#_Toc9981498)

[IV. Các hàm và toán tử 7](#_Toc9981499)

[V. Các công nghệ và thư viện hỗ trợ khác 8](#_Toc9981500)

[1. Jsoup là gì? 8](#_Toc9981501)

[2. ví dụ 8](#_Toc9981502)

[3. Thư viện JSOUP 9](#_Toc9981503)

[CHƯƠNG IV Phân tích 9](#_Toc9981504)

[I . Một ví dụ tổng quát của bài toán thu thập dữ liệu 9](#_Toc9981505)

[V THIẾT KẾ 10](#_Toc9981506)

[I . Crawler 10](#_Toc9981507)

[II. Lấy thông tin trên một trang web và xử lí nó. 11](#_Toc9981508)

[1. Lấy thông tin trên một trang web cụ thể 11](#_Toc9981509)

[2. Tự động thu thập dữ liệu 13](#_Toc9981510)

[3. Xử lí dữ liệu 13](#_Toc9981511)

# CHƯƠNG I : GIỚI THIỆU

## I. Mở đầu

Với sự phát triển mạnh mẽ của Internet trong những năm gần đây, đặc biệt là sự bùng nổ của World Wide Web. Những người dùng bình thường có thể tìm kiếm thông tin trên web một cách dễ dàng nhờ vào các máy tìm kiếm (search engine), họ có thể đánh dấu (bookmark) các trang kết quả hoặc sao chép nội dung họ muốn vào một nguồn nào đó để tham khảo về sau.

Nhằm phục vụ mục đích khảo sát thị trường, nhu cầu tổng hợp tin tức từ các website tin tức để xây dựng các trang web thông tin tổng hợp, nhu cầu thu thập thông tin về các doanh nghiệp thuộc một ngành nào đó trên website danh bạ doanh nghiệp để gửi email quảng cáo, tiếp thị, v.v… Chính những ví dụ thực tế như trên đã nảy sinh ra nhu cầu: cần phải có một phương thức hoặc công cụ nào đó có khả năng tìm kiếm, trích xuất thông tin trên web và lưu trữ lại thông tin đó theo ý muốn của con người, một cách tự động và hiệu quả.

### 1. Yêu cầu và mục tiêu của đề tài

### 1.1 Yêu cầu

Đề tài: Tìm hiểu ứng dụng Crawl lấy dữ liệu và tin tức từ các website

Con người có nhu cầu thu thập một khối lượng lớn dữ liệu từ các website để phục vụ cho mục đích của mình (chẳng hạn thu thập dữ liệu về các nhà cung cấp sản phẩm thời trang trên website Trang Vàng Việt Nam để gửi email quảng cáo). Lượng dữ liệu này nếu thu thập bằng tay thì phải mất rất nhiều thời gian và công sức. Do đó, trong đề tài này, sinh viên sẽ tìm hiểu về các kỹ thuật thu thập dữ liệu trên web một cách tự động và hiện thực một giải pháp cụ thể để thu thập chúng.

Các nhiệm vụ:

* Tìm hiểu về các công nghệ hỗ trợ thu thập dữ liệu trong ngôn ngữ Java
* Áp dụng công nghệ để lấy dữ liệu từ một vài website cụ thể
* Thực hiện xử lí phân tích dữ liệu thô đã lấy được.

### 1.2 Mục tiêu

Lấy được thông tin từ mạng internet và xử lí để lấy thông tin cần thiết.

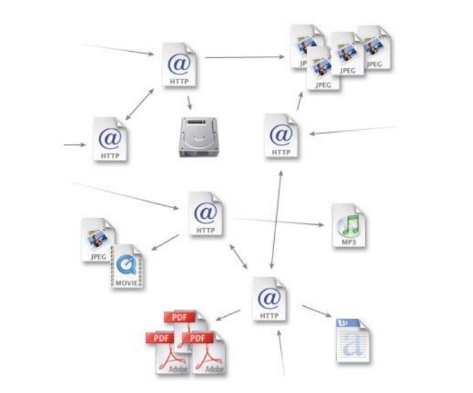
# CHƯƠNG II. TÌM HIỂU VÀ PHÂN TÍCH SƠ BỘ

Để có thể đi sâu vào các vấn đề đề tài đặt ra, trước hết chúng ta cùng tìm hiểu một số kiến thức khái quát để có cái nhìn tổng quan về bài toán

## I . Tìm hiểu các kiến thức tổng quan

### 1. World Wide Web

World Wide Web (WWW, hay gọi tắt là Web) là một ứng dụng phổ biến và phát triển mạnh mẽ nhất của Internet hiện nay. World Wide Web là một mạng lưới bao gồm các tài liệu siêu văn bản (hypertext) được đặt trên các máy tính nằm trong mạng Internet.

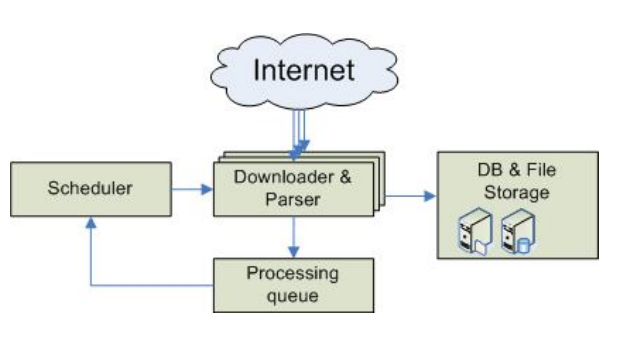


Hình 1.1 World wide web

Nội dung các trang web chủ yếu được viết bằng ngôn ngữ HTML hoặc XHTML. Khi muốn truy cập một trang web, trình duyệt web sẽ gửi yêu cầu đến máy chủ (web server) chứa trang web đó.

### 2. Web Crawler

Web crawler có chức năng lấy thông tin từ website , trích xuất ra những thông tin người sử dụng cần, đồng thời cũng tìm những link có trong trang web đó và tự động truy cập vào những link đó. Các tên gọi khác của crawler là robot, bot, spider, worm, ant. Nhưng gần đây tên gọi crawler là thông dụng nhất.



Hình 1.2 hoạt động của một web crawler đơn giản

### 3. Web Scraper

Các trang web chủ yếu được viết bằng các ngôn ngữ đánh dấu như HTML, XHTML và được nhắm đến đối tượng sử dụng là con người chứ không phải máy tính. Các trang web lại chứa đựng nhiều thông tin có ích mà con người có thể muốn thu thập và lưu trữ lại, chính vì thế mà các web scraper được ra đời.

## II. Phân loại Web

World Wide Web có thể được phân loại thành hai loại: các trang web tĩnh và các trang web động. Sriram và Hector đưa ra định nghĩa sau về trang web động:

Một trang P được gọi là động nếu như một phần hoặc tất cả nội dung của nó được sinh ra tại thời điểm chạy (tức là sau khi yêu cầu của máy khách được máy chủ nhận) bởi một chương trình thực thi nằm trên máy chủ hoặc máy khách.

### 1. Phân loại dựa vào sự thể hiện của tính động

**Theo thời gian (temporal dynamism):** đây là những trang web mà nội dung của chúng có thể được thay đổi, cập nhật theo thời gian.

**Theo máy khách (client-based dynamism):** những trang web có khả năng tùy biến theo người sử dụng (client) sẽ được xếp vào mục này.

Ví dụ một trang tin tức tổng hợp có khả năng chọn lọc các tin tức khác nhau, tùy thuộc vào sở thích của người sử dụng đang đăng.

**Theo truy vấn (input dynamism):** đây là những trang có nội dung phụ thuộc vào truy vấn của người sử dụng. Một ví dụ điển hình là những trang có các mẫu nhập liệu (form), chẳng hạn một trường nhập liệu để tìm kiếm một hoặc nhiều món hàng trong cơ sở dữ liệu của trang web.

### 2. Phân loại dựa vào cách thức tạo ra tính động

Thực thi các chương trình nằm trên máy chủ (server-side programs): Trong kỹ thuật này, một chương trình sẽ được thực thi trên máy chủ và sinh ra toàn bộ nội dung HTML của trang web, sau đó được gửi đến máy khách yêu cầu.

Ví dụ tiêu biểu cho loại này là CGI hoặc Java Servlet. Những chương trình phía máy chủ này cũng thường được dùng để xử lý các truy vấn từ người sử dụng.

**Nhúng mã với sự thực thi phía máy chủ** (embedded code with server-side execution): Khi một yêu cầu được nhận, các đoạn mã nhúng này sẽ được thực thi trên máy chủ và sẽ sinh ra các đoạn code HTML thay thế cho chúng. Kỹ thuật này khác kỹ thuật trên ở chỗ không phải toàn bộ mà chỉ một phần nội dung HTML được sinh động. Các mã nhúng này có thể là PHP, Java Scriplet, ASP hoặc các mã server-side khác.

**Nhúng mã với sự thực thi phía máy khách** (có thể có cả sự thực thi ở phía máy chủ) (embedded code with client-side execution): Với ký thuật này thì các đoạn mã được nhúng không chạy trên server mà chúng được máy khách tải về và được thực thi trên máy khách. Nếu mã chỉ chạy trên máy khách thì đó có thể là JavaScript, Java Applet hoặc Flash. Nếu vừa có sự thực thi ở máy khách và cả ở máy chủ thì đó chính là Ajax, trong trường hợp này mã clien-side có thể trao đổi thông tin với server thông qua một đối tượng đặc biệt là XMLHttpRequest để có thể thay đổi nội dung trang web một cách động mà không cần load trang web mới.

**2.1 Khả năng của các web crawler**

Các crawler này thường phải bỏ qua các trang có tính động theo máy khách (client-based dynamism) và theo truy vấn (input dynamism). Cũng đã có một vài công trình nghiên cứu xây dựng các hidden-web crawler như HiWE .

# CHƯƠNG III. CÁC KIẾN THỨC NỀN TẢNG

## I . HTML

**HTML** là chữ viết tắt của **Hypertext Markup Language**. Nó giúp người dùng tạo và cấu trúc các thành phần trong trang web hoặc ứng dụng, phân chia các đoạn văn, heading, links, blockquotes, ..

## II . XML

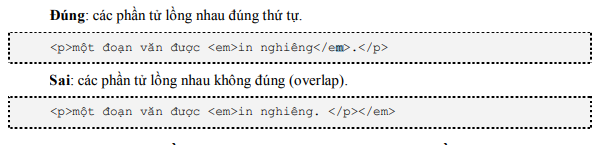
XML (Extensible Markup Language) là ngôn ngữ đánh dấu mở rộng. Đây là một dạng ngôn ngữ đánh dấu, có chức năng truyền dữ liệu và mô tả nhiều loại dữ liệu khác nhau.

  
hình 3.1 XML

## III . XHTML

XHTML đòi hỏi các tài liệu viết bằng XHTML phải có tính chất well-formed, có nghĩa là tất cả các thẻ trong tài liệu đều phải có thẻ đóng tương ứng (ngoại trừ các thẻ rỗng được trình bày cụ thể ở phần sau) và các phần tử phải lồng nhau một cách đúng thứ tự (thẻ nào mở trước thì phải đóng sau).

Ví dụ:

**

Hình 3.2. Tính chuẩn của một tài liệu XML

## IV. Các hàm và toán tử

XPath 1.0 định nghĩa 4 kiểu dữ liệu: kiểu tập node (node-set), kiểu chuỗi ký tự (string), kiểu số (number) và kiểu logic (boolean).

Các toán tử phổ biến được dùng trong XPath: • Toán tử hội “|“ để tìm hội của hai tập node.

• Toán tử logic “and” và “or” (và một hàm not(boolean) để phủ định một biểu thức boolean).

• Các toán tử cộng trừ nhân chi “+”, “-“, “\*”, “div” và “mod”. • Các toán tử so sánh “=”, “!=”, “”, “<=”, “>=”. Các hàm phổ biến trong thự viện hàm của XPath 1.0:

• position(): trả về giá trị kiểu số thể hiện vị trí của node trong một chuỗi các node.

• count(node-set): trả về số node có trong node-set.

• string(object?): chuyển bất kỳ kiểu dữ liệu nào thành kiểu chuỗi. Nếu tham số là kiểu node-set thì hàm này trả về giá trị chuỗi của node đầu tiên trong node-set (theo thứ tự xuất hiện trong tài liệu).

• contains(s1, s2): trả về true nếu s1 chứa s2.

• true(), false(): trả về các giá trị true, false tương ứng.

Một số ví dụ:

*//item[@price > 2\*@discount]*

trả về các node item có thuộc tính price lớn hơn hai lần giá trị thuộc tính discount của nó;

*//product[contains(comment, ‘bad’)]*

trả về các node product có node con comment chứa từ ‘bad’

## V. Các công nghệ và thư viện hỗ trợ khác

### 1. Jsoup là gì?

Jsoup là một thư viện được sử dụng để phân tích tài liệu HTML. Jsoup cung cấp các API dùng để lấy dữ liệu và thao tác dữ liệu từ URL hoặc từ tập tin HTML.



Hình 3.2 biểu tượng JSOUP

### 2. ví dụ

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13 | import java.io.IOException;  import org.jsoup.Jsoup;  import org.jsoup.nodes.Document;    public class HelloJsoup {       public static void main( String[] args ) throws IOException{         Document doc = Jsoup.connect("http://baomoi.com").get();         String title = doc.title();         System.out.println("Title : " + title);     }    } |

### 3. Thư viện JSOUP

Có thể sử dụng Maven hoặc download thư viện Jsoup dưới dạng file jar.

Với maven

<!-- <http://mvnrepository.com/artifact/org.jsoup/jsoup> -->

<dependency>

   <groupId>org.jsoup</groupId>

   <artifactId>jsoup</artifactId>

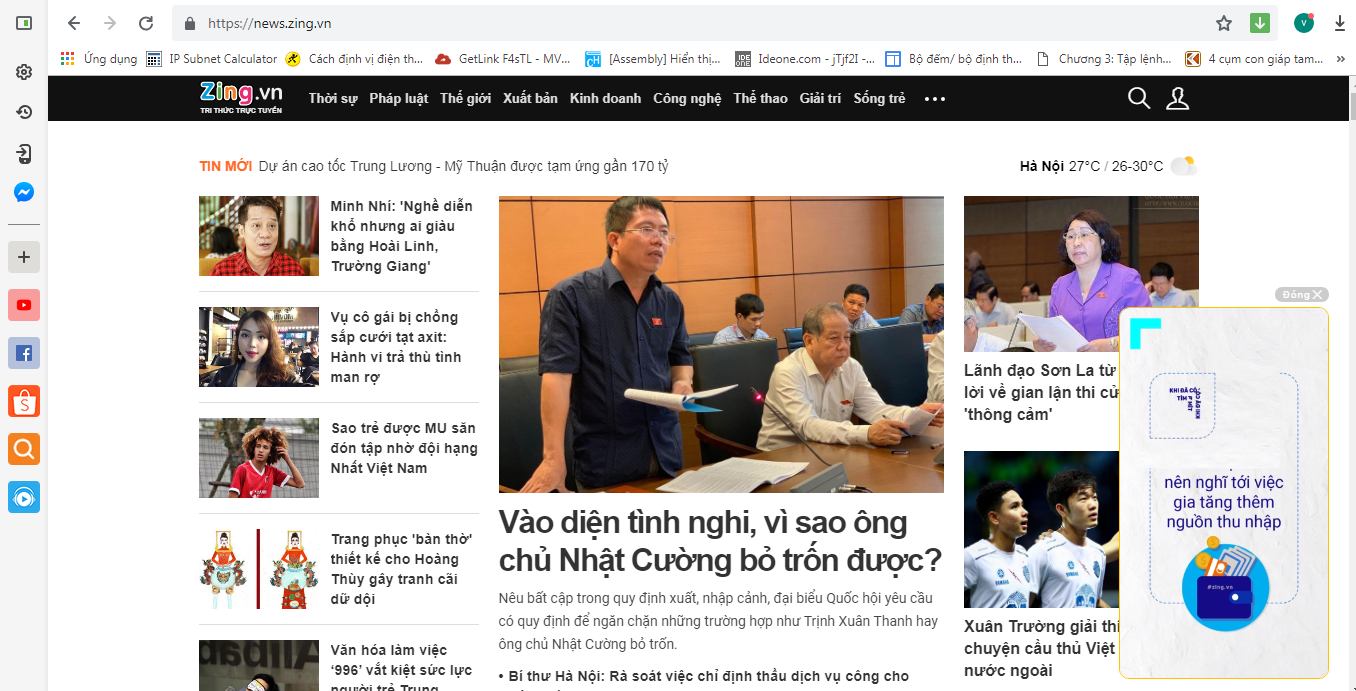
   <version>1.8.3</version>

</dependency>

# CHƯƠNG IV Phân tích

## I . Một ví dụ tổng quát của bài toán thu thập dữ liệu

Các trang chi tiết này chính là các trang cấp thấp nhất, nội dung của trang này sẽ chứa các thông tin mà người sử dụng website quan tâm như: tên doanh nghiệp, mô tả về doanh nghiệp, địa chỉ liên lạc, số điện thoại, địa chỉ email v.v…

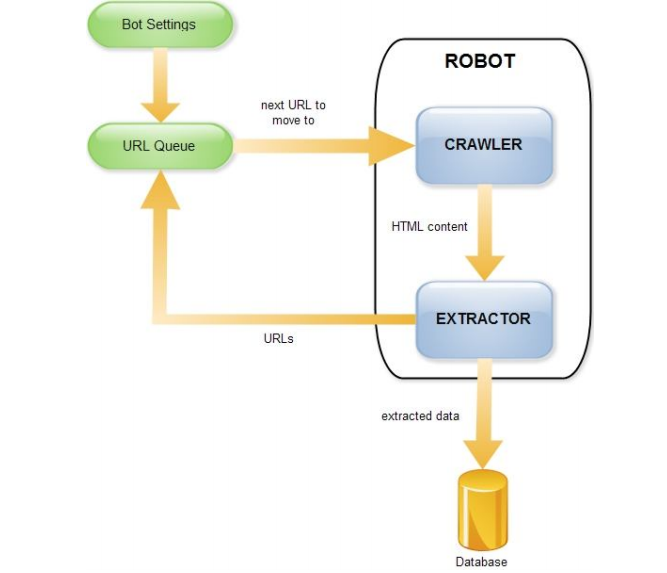


Hình 4.1 một trang web của zing.vn

# V THIẾT KẾ

## I . Crawler

Crawler đảm nhận trách nhiệm quản lý các URL đã ghé thăm cũng như cần được ghé thăm tiếp theo, gửi yêu cầu đến máy chủ, tải nội dung trang web về dưới dạng HTML text. Các URL cần được ghé thăm được lưu trong cơ sở dữ liệu của Bot, sẽ được Crawler đọc lên, cùng với các thông tin hỗ trợ khác thể hiện cách thức chuyển trang (get, post, parameters, javascript …) để Crawler có thể lấy được đúng nội dung của các trang web tiếp theo.

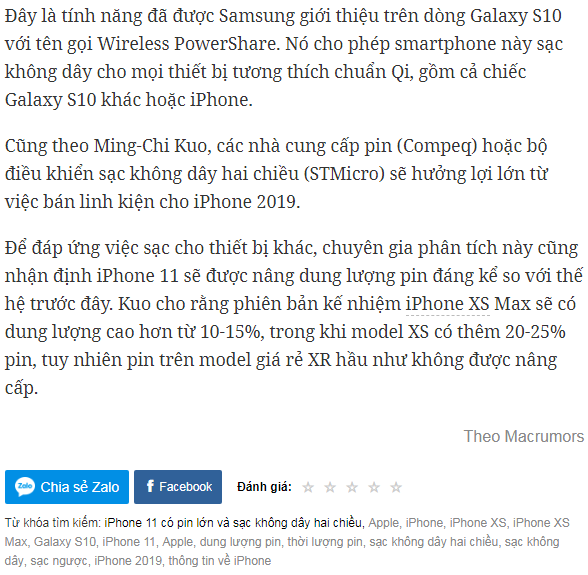
****

Hình 5.1 hoạt động của robot

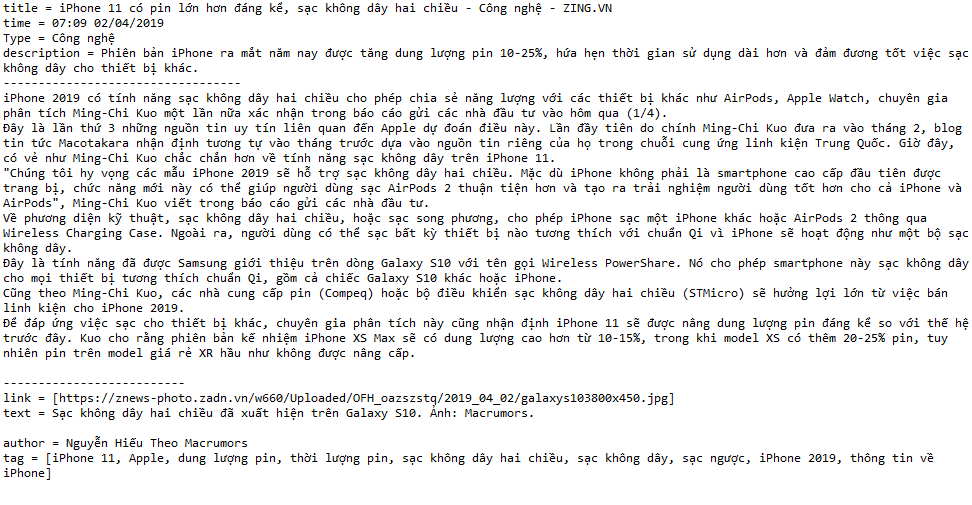
## II. Lấy thông tin trên một trang web và xử lí nó.

### Lấy thông tin trên một trang web cụ thể

Từ một trang web zing cụ thể



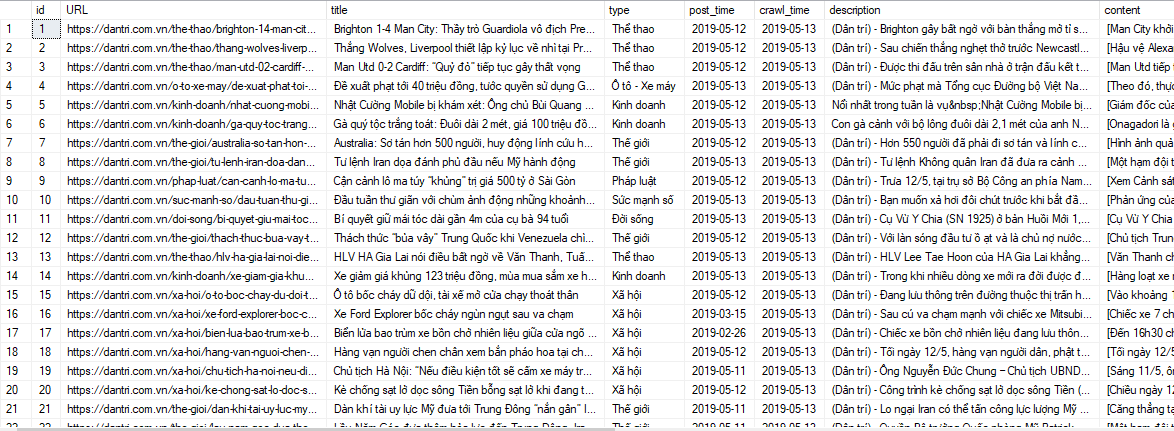
Chúng ta có thể lấy được các dữ liệu như sau:



Sau khi đã có thể lấy dữ liệu của 1 trang, chúng ta có thể viết chương trình để nó tự động lấy hết dữ liệu của một hoặc một vài trang báo nào đó.

### Tự động thu thập dữ liệu

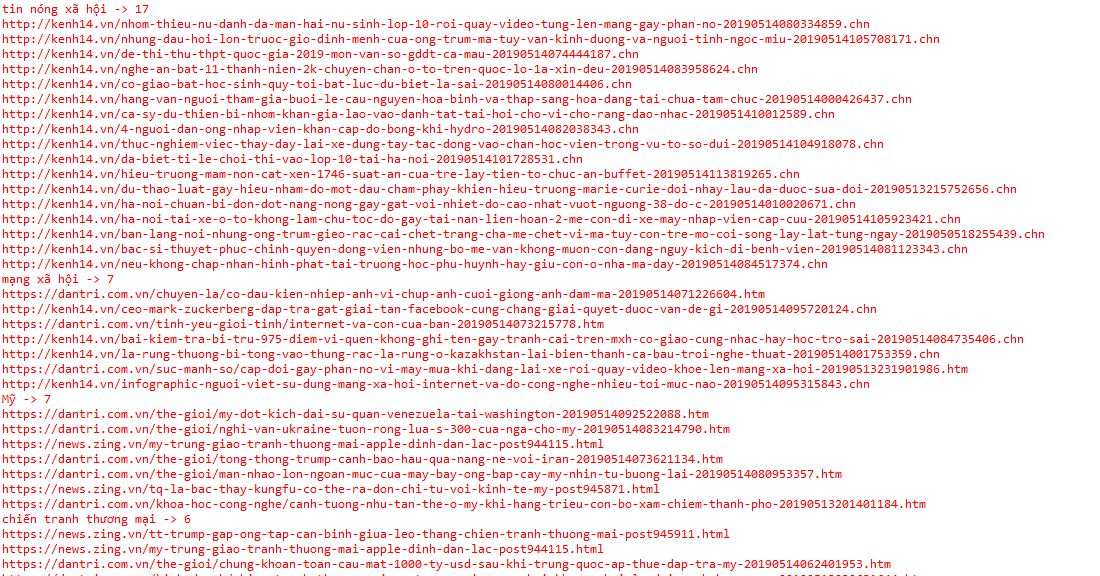
Việc lấy tự động được thực hiện và có kết quả



Các dữ liệu sau khi lấy được sẽ được đưa vào database, sau đó chúng ta có thể xử lí dử liệu theo ý của bản thân.

### Xử lí dữ liệu

Sau đó, dựa trên dữ liệu đã lấy, chúng ta thực hiện xử lí dữ liệu, thống kê các bài báo, các nhãn tag hot trong ngày / tuần / tháng,…



Thống kê dữ liệu theo tuần

Ngoài những thống kê đã được xử lí ở đây, chúng ta có thể xử lí những thống kê khác như: lấy các tag hot trên mỗi lĩnh vực như Kinh tế, Thể Thao,…

Việc lấy dữ liệu có thể giúp cho chúng ta rất nhiều. Hiện nay, dữ liệu được ví như vàng. Nếu có càng nhiều dữ liệu thì chúng ta có thêm nhiều thông tin. Việc crawl và xử lí dữ liệu sẽ là một bước tiến lớn giúp cho nhóm về sau.