**TIÊU ĐỀ CHƯƠNG**

**TIÊU ĐỀ MỤC**

*TIỂU MỤC*

TIỂU MỤC CON

**CẤU TRÚC BÁO CÁO**

- Tổng quan / Giới thiệu đề tài

- Cơ sở lý thuyết / Thực trạng vấn đề / Giới thiệu công nghệ

- Phân tích thiết kế

- Cài đặt & Đánh giá kết quả

- Kết luận

**Hợp Đồng Win - Win**

**0.1 Lê Khôi Đệ**

*0.1.1 Kết quả mong muốn*

0.1.1.1 Kết quả đồ án tổng hợp, bản thân mình muốn nó như thế nào chi tiết cụ thể?

* Đạt được điểm B+ tương đương khá tốt
* Tạo được phần mềm đề ghi điểm nhà tuyển dụng, khoe cho anh trai - khoe facebook để mọi người biết tôi là ai nếu cần giúp đỡ về công nghệ, dấu ấn cuộc đời về kỹ sư công nghệ thông tin của tôi,.

0.1.1.2 Để đạt được kết quả trên thì bản thân mình phải rèn luyện như thế nào?

* Khi gặp vấn đề khúc mắc thì phải chủ động hỏi giảng viên hướng dẫn. Rèn luyện tính chủ động gạt bỏ sự ngại ngùng
* Hạn chế thời gian chơi game và xem youtube. Quản lý thời gian tập trung cho điều quan trọng nhất

*0.1.2 Đường lối thực hiện*

Vai trò của bản thân muốn làm gì sau khi đã đi qua 2 đồ án trước?

* Phân tích và lên kế hoạch

*0.1.3 Tài nguyên*

Những người bạn có thể giúp mình đạt được kết quả là những ai?

* Kiddy Thanh tên facebook học khá tiếng Anh có thể giúp đỡ dịch thuật những tài liệu tiếng Anh
* Peter Truong tên Facebook hiện là học vẽ đồ họa chuyên về Web có thể giúp vẽ những giao diện Web đẹp
* Anh Thông học môn Lập Trình Web đã từng chỉ về cách lấy dữ liệu một trang Web là sách để đọc

*0.1.4 Báo cáo tiến độ*

0.1.4.1 Nhu cầu tâm lý khi có vấn đề (lúc lười biếng, mệt mỏi, thiếu tiền, bận bịu) thì bản thân sẽ làm như thế nào để không ảnh hưởng đến kết quả đồ án?

* Thì sẽ khá im lặng mong đối phương thông cảm và tế nhị
* Một người dễ bị chi phối cảm xúc và bản thân sẽ nhìn nhận, điềm tĩnh tránh xảy ra xung đột.

0.1.4.2 Và muốn người bạn cùng nhóm giúp đỡ như thế nào (đá đít hay dẫn đi ăn ngon)

* Khi mệt mỏi căng thẳng mong người bạn cùng nhóm đừng làm Stress thêm bằng cách giao thêm công việc

0.1.4.3 Kết quả công việc mỗi tuần nộp một bản chính cho cô Thơ và một bản phụ qua ZALO

* Oke thiếu một bản nào thì nộp phạt 100k

*0.1.5 Cam kết hậu quả*

0.1.5.1 Đi trễ họp thì phạt như thế nào?

* Phạt 20k tương đương một tô hủ tiếu

0.1.5.2 Trễ Deadline thì phạt như thế nào? (Deadline là những công việc từ buổi họp trước đến buổi họp tiếp theo phải có kết quả)

* Phạt 20k tương đương một tô hủ tiếu.

**0.2 Nguyễn Bá Tâm**

*0.1.1 Kết quả mong muốn*

0.1.1.1 Kết quả đồ án tổng hợp, bản thân mình muốn nó như thế nào chi tiết cụ thể?

* Mong muốn bản thân: Hoàn thành đồ án tổng hợp, không bị nợ và kết thúc chương trình học đại học

0.1.1.2 Để đạt được kết quả trên thì bản thân mình phải rèn luyện như thế nào?

* Dành hầu hết thời gian vào việc tìm hiểu và thực hiện đồ án
* Không để các yếu tố bên ngoài tác động vào quá trình thực hiện đồ án tổng hợp

0.1.2 Đường lối thực hiện

Vai trò của bản thân muốn làm gì sau khi đã đi qua 2 đồ án trước?

* Đưa ra ý kiến bản thân và thực hiện các công việc được giao bởi leader.

0.1.3 Tài nguyên

Những người bạn có thể giúp mình đạt được kết quả là những ai?

\_ Những người bạn có hiểu biết về công nghệ thực hiện đồ án

\_ Giảng viên hướng dẫn đồ án tổng hợp

\_ Những anh chị khóa trước đã thực hiện đồ án tổng hợp hoặc đã đi làm có kiến thức hiểu biết về công nghệ làm đồ án.

*0.1.4 Báo cáo tiến độ*

0.1.4.1 Nhu cầu tâm lý khi có vấn đề (lúc lười biếng, mệt mỏi, thiếu tiền, bận bịu) thì bản thân sẽ làm như thế nào để không ảnh hưởng đến kết quả đồ án?

* Trình bày rõ ràng các lý do ảnh hưởng đến tâm lý bản thân. Báo cho leader biết để cùng tìm cách giải quyết không để ảnh hưởng đến tiến độ thực hiện đồ án tổng hợp.

0.1.4.2 Và muốn người bạn cùng nhóm giúp đỡ như thế nào (đá đít hay dẫn đi ăn ngon)

- Do kiến thức còn hạn hẹp nên mong bạn cùng nhóm hạn chế gây thêm áp lực.

0.1.4.3 Kết quả công việc mỗi tuần nộp một bản chính cho cô Thơ và một bản phụ qua ZALO

Oke thiếu một bản nào thì nộp phạt 100k

*0.1.5 Cam kết hậu quả*

0.1.5.1 Đi trễ họp thì phạt như thế nào?

Phạt 20k tương đương một tô hủ tiếu

0.1.5.2 Trễ Deadline thì phạt như thế nào? (Deadline là những công việc từ buổi họp trước đến buổi họp tiếp theo phải có kết quả)

Phạt 20k tương đương một tô hủ tiếu

* 1. **WBS (Cấu trúc phân rã công việc)**

*1.1.1 Xác định (10/02/2020 - 29/02/2020)*

1.1.1.1 Xác định ý tưởng thực hiện đề tài (10/02/2020 - 29/02/2020)

Kết quả: Giới thiệu về đề tài.docx

Chịu trách nhiệm: Lê Khôi Đệ.

1.1.1.2 Xác định thực trạng vấn đề / mục đích lý do về ý tưởng thực hiện đề tài (10/02/2020 - 29/02/2020)

Kết quả: Thực trạng vấn đề.docx

Chịu trách nhiệm: Lê Khôi Đệ

1.1.1.3 Xác định yêu cầu về giao diện (24/02/2020 - 28/02/2020)

Kết quả: Yêu cầu giao diện.docx

Chịu trách nhiệm: Lê Khôi Đệ

1.1.1.4 Xác định công nghệ và ngôn ngữ lập trình sử dụng thực hiện đề tài (24/02/2020 - 28/02/2020)

Kết quả: Giới thiệu công nghệ.docx

Chịu trách nhiệm: Nguyễn Bá Tâm - Lê Khôi Đệ

1.1.1.5 Xác định thời gian bắt đầu và kết thúc của đề tài

Thời gian thực hiện đồ án: 10/02/2020 - 6/06/2020

Thời gian nộp báo cáo và Poster đề tài: 8/06/2020 - 12/06/2020

Thời gian báo cáo trước hội đồng: 20/06/2020 - 27/06/2020

Chịu trách nhiệm: Khoa Công Nghệ Thông Tin

*1.1.2 Phân tích và thiết kế (1/03/2020 - 04/04/2020)*

1.1.2.1 Phân tích chức năng (01/03/2020 - 04/04/2020)

Crawl dữ liệu (01/03/2020 - 04/04/2020)

Kết quả: Tài liệu phân tích nghiên cứu 1.docx

Chịu trách nhiệm: Nguyễn Bá Tâm

Dữ liệu khi cào về được Import vào Database(01/03/2020 - 04/04/2020)

Kết quả: Tài liệu phân tích nghiên cứu 2.docx

Chịu trách nhiệm: Lê Khôi Đệ

1.1.2.2 Phân tích thiết kế giao diện (22/03/2020 - 04/04/2020)

Giao diện trang chủ người dùng (22/03/2020 - 28/03/2020)

Kết quả: Tài liệu hình ảnh và kèm lời chú thích 1.docx

Giao diện người dùng xem chi tiết một sự kiện (29/03/2020 - 04/04/2020)

Kết quả: Tài liệu hình ảnh và kèm lời chú thích 2.docx

*1.1.3 Thực hiện (05/04/2020 - 23/05/2020)*

1.1.3.1 Thực hiện hệ thống (05/04/2020 - 11/04/2020)

Cập nhật cơ sở dữ liệu Hutech 2020 trên SQL Server

Kết quả: Tài liệu hình ảnh và kèm lời giải thích.docx

Chịu trách nhiệm: Lê Khôi Đệ

1.1.3.2 Thực hiện chức năng (05/04/2020 - 23/05/2020)

Crawl dữ liệu (05/04/2020 - 23/05/2020)

Đầu vào: Tài liệu phân tích nghiên cứu 1.docx

Kết quả: Tài liệu hình ảnh và kèm lời giải thích 1.docx

Chịu trách nhiệm: Nguyễn Bá Tâm

Dữ liệu khi cào về được Import vào Database (03/05/2020 - 23/05/2020)

Đầu vào: Tài liệu phân tích nghiên cứu 2.docx

Kết quả: Tài liệu hình ảnh và kèm lời giải thích 2.docx

Chịu trách nhiệm: Lê Khôi Đệ

1.1.3.3 Thực hiện giao diện lên hệ thống(12/04/2020 - 02/05/2020)

Kết quả: Tài liệu hình ảnh và kèm lời giải thích.docx

1.1.4 *Kiểm thử (24/05/2020 - 30/05/2020)*

1.1.4.1 Kiểm thử chức năng (24/05/2020 - 30/05/2020)

Kết quả: Tài liệu đánh giá kết quả.docx

Chịu trách nhiệm: Nguyễn Bá Tâm - Lê Khôi Đệ

*1.1.5 Kết luận (31/05/2020 - 27/06/2020)*

1.1.5.1 Viết và chỉnh sửa báo cáo (31/05/2020 - 12/06/2020)

Chịu trách nhiệm: Lê Khôi Đệ

1.1.5.2 Làm Slide thuyết trình (13/06/2020 - 20/06/2020)

Chịu trách nhiệm: Nguyễn Bá Tâm - Lê Khôi Đệ

1.1.5.3 Thuyết trình (20/06/2020 - 27/06/2020)

Chịu trách nhiệm: Nguyễn Bá Tâm

**1.2 Yêu cầu giao diện**

Yêu cầu tổng thể giao diện người dùng:

1. Điều hướng đơn giản người dùng phải tìm thấy nội dung trong vài giây
2. Bố cục đơn giản, hình ảnh sự kiện đóng khung làm nổi bật hình ảnh bên trong, các khung có trúc gọn gàng xung quanh tạo cảm giác sắp xếp có thứ tự
3. Chân trang nên đặt những thông tin rõ ràng
4. Logo tránh móp mép, hình ảnh sự kiện rõ ràng - tránh bị nhòe
5. Tập trung vào trải nghiệm người dùng
6. Đáp ứng mọi thiết bị di động

Cách làm WBS

Danh sach chi tiết những gì cần hoàn thành một dự án

- Sẽ xác định được các bước để hoàn thành dự án

- Xác định trách nhiệm giữa các cá nhân

- Có 3 cách tiếp cận:

+ Tiếp cận tương tự : Xem các WBS của các dự án tương tự và sửa đổi cho phù hợp

+ Từ trên xuống: Bắt đầu với thành phần lớn nhất, sau đó chia nhỏ dần

+ Từ dưới lên: Bắt đầu từ công việc chi tiết, kết hợp thành công việc lớn

- Những yếu tố trong bảng công việc

+ Danh sách các sản phẩm

+ Ước lượng thời gian

+ Lịch trình thực hiện

+Phân bố lực lượng, tài nguyên

+ Tính chi phí

+ Kiểm soát rủi ro

- Cấu trúc bảng công việc : DSSP, DSCV

+ DSSP: bằng danh từ

+ DSCV: động từ và một bổ ngữ

- Các bước xây dựng WBS : viết ra sản phẩm chung hết, tạo danh sách sản phẩm, tạo danh sách công việc, đánh mã cho mỗi sản phẩm - công việc, xét duyệt lại WBS

- Lưu ý:

+ Từng công việc trong WBS được chi tiết tới mức thấp nhất (<80 giờ)

+ Đạt được sự đồng thuận giữa các bên

+ Chuẩn bị bản thảo của WBS, gửi cho mọi người đọc trước

+ Họp thảo luận, đi đến nhất trí và ký

+ Thường xuyên cập nhật

B1: Thu thập thông tin từ các thành viên => tách nhỏ các nhiện vụ chi tiết hơn

B2: Đối với từng nhiệm vụ:

- Xác định công việc và kết quả mong muốn đạt được

- Liệt kê danh sách cá nhân chịu trách nhiệm

- Xây dựng sơ bộ thời gian cần thiết để thực hiện từng nhiệm vụ trong WBS

B3: Xem xét lại WBS, ngân sách và thời gian, đánh giá với các thành viên sẽ chịu trách nhiệm thực hiện công việc đó => kiểm tra mức độ chính xác của WBS

B4: Sắp xếp các nhiệm vụ theo thứ bậc

B5: Chi phí

B6,7: Khỏi cần

B8: Kiểm tra nguồn lực theo mức độ công việc

B9: Khỏi

Cơ sở lý thuyết

- Lưu nhật kí tiến độ công việc bao gồm làm được và chưa làm được. Ghi nhận xét

- Lưu thông tin cụ thể về tài liệu đã đọc để sau này nghiên cứu thêm và trích dẫn theo đúng quy định

- Tóm tắt khi đọc tài liệu đó, ghi chú lại nhận xét, đánh dấu phần đọc được cẩn thận để ghi lại vào báo cáo

- Tham khảo để đánh giá nghiên cứu và vấn đề

- Xác định chủ đề quan tâm: là nội dung chính

1.1 Tra cứu và cập nhật dữ liệu các thông tin Hutech năm 2020

1.2 Tìm hiểu Crawl dữ liệu

* Luận đề: Là vấn đề cần giải quyết. Nói theo cách logic thì đó là phán đoán cần chứng minh.
* Luận điểm: Là giả thuyết, nhận định, ý kiến sau quá trình luận chứng của người viết về vấn đề được nếu ra
* Luận cứ: Là bằng chứng để chứng minh
* Luận chứng: Luận chứng là quá trình phối hợp, tổ chức, thực hiện, phản biện để đưa các lý lẽ và cuối cùng chứng minh cho luận điểm hoặc tạo ra luận điểm. Luận chứng phải chặt chẽ, logic, hài hòa, unbiased (không phân cực) hay dùng phương pháp luận cho thuyết phục.

TUẦN 2

**BÁO CÁO TUẦN 2**

**2.1 Dữ liệu khi cào về được Import vào Database**

<https://freetuts.net/huong-dan-crawl-du-lieu-trang-chonbaihat-bang-ajax-+-php-1666.html>

Điều quan trọng rút ra được:Trước khi Crawl một trang Web thì phải soi toàn bộ cấu trúc HTML để đưa ra phương án nghĩa là mỗi trang phải code khác nhau vì cấu trúc khác nhau

Cảm nhận: Hơi sai về công nghệ muốn sử dụng, nhưng nó tập trung vào những gì theo hướng nghiên cứu. Dữ liệu khi cào về được Import vào Database sẽ được 1 trang “save.php” đảm nhận

Tóm tắt quá trình họ làm:

- Họ dùng CSS Selector kết hợp jQuery Selector để bóc nội dung của từng bài hát

- Họ dùng 3 trang như sau

+ index.php: đây là trang chính xử lý bóc tách dữ liệu

+ getpage.php: do JS không thể lấy content của một trang web online được nên trang này sẽ có nhiệm vụ trả về mã HTML của một trang Web

+ save.php: là trang lưu bài hát vào CSDL, trang này mình sẽ không code mà bạn phải tự code nhé.

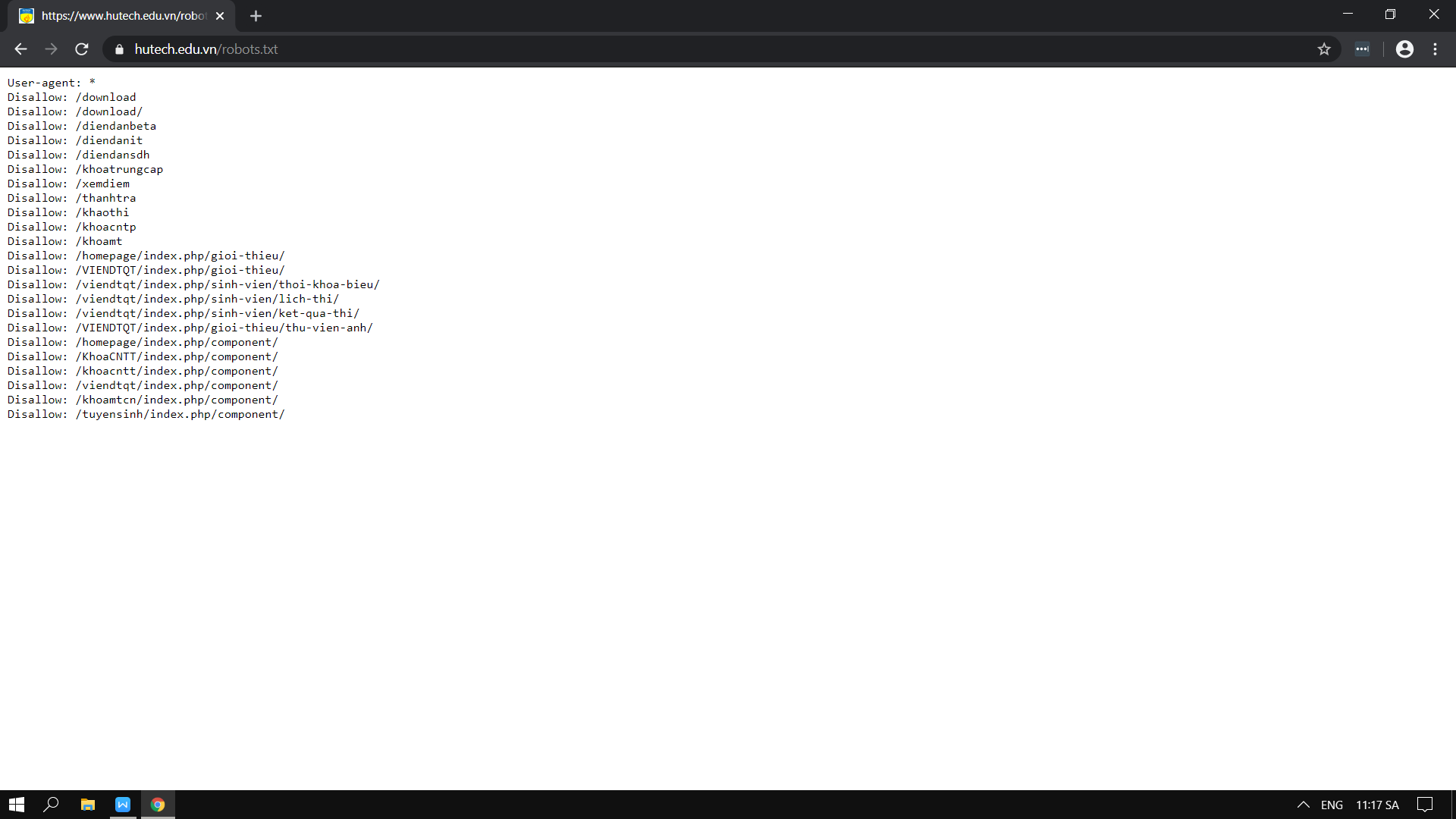
- File getpage.php có một đoạn code trả về nội dung của một trang web dựa vào tham số URL

- File save.php in ra giá trị của biến POST, dựa vào giá trị lưu vào CSDL

<http://viet.jnlp.org/tai-nguyen-ngon-ngu-tieng-viet/ke-hoach-xay-dung-tu-dong-corpus-tu-nguon-web/crawler>

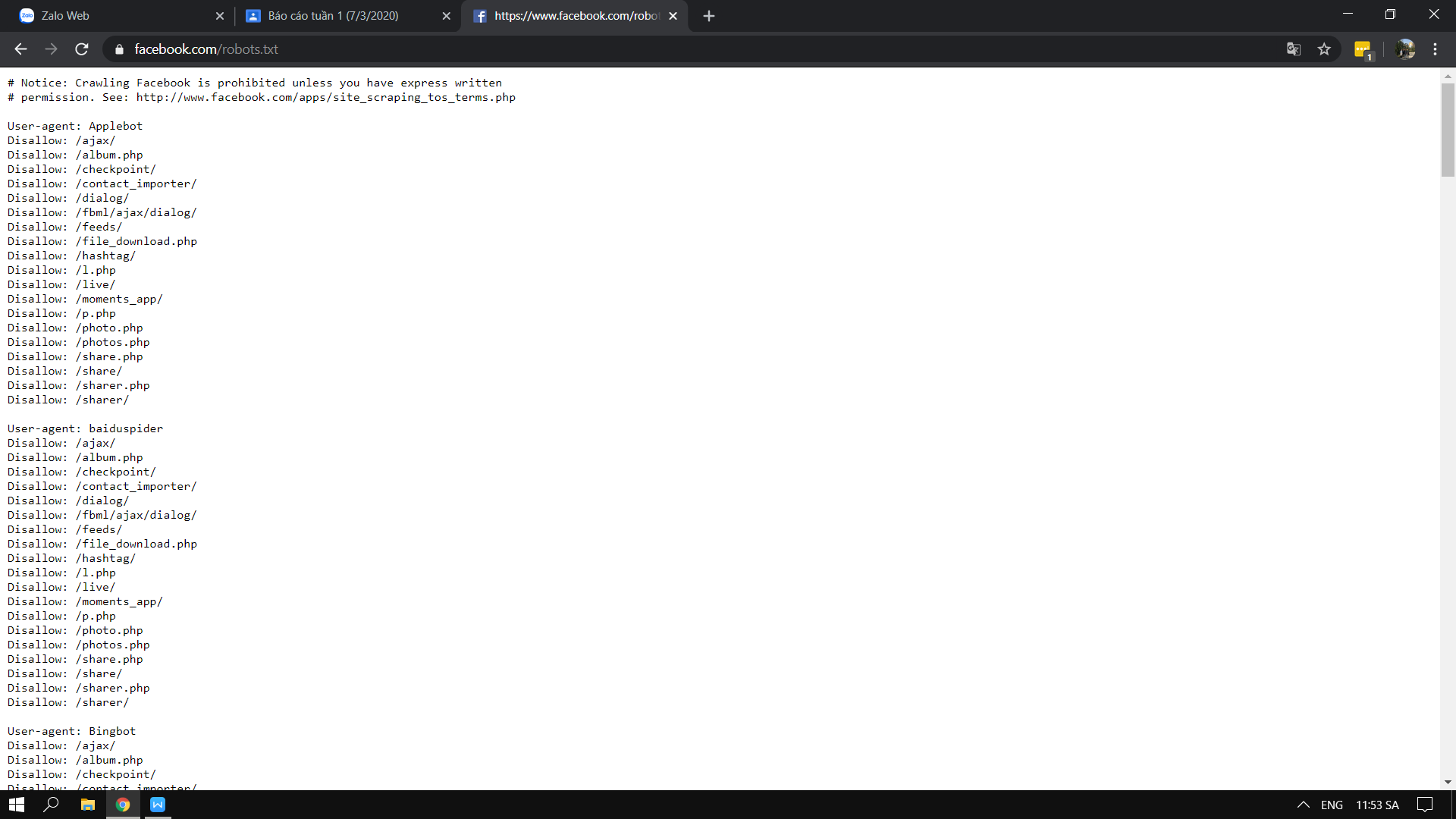
Những điều quan trọng rút ra được:

1. Trường hợp nếu Server không phản hồi khi gửi Request thì chương trình sẽ bị đóng băng một rủi ro phải khắc phục
2. Nếu gửi quá nhiều Request thì Website có thể bị nghẽn, nặng hơn là Dos một rủi ro phải khắc phục
3. Phải đánh dấu những Link đã được xử lý tránh tình trạng vòng lặp một rủi ro phải khắc phục
4. Có những link khác nhau từ một trang Web phải xử lý code khác nhau một kinh nghiệm phải học hỏi
5. Một trang Website thường có File “robots.txt” để giới hạn Crawler vậy nên phải ưu tiên xem trước những Link bị cấm Crawler



Hình 2.1 robots.txt từ Hutech

1. Theo hình trên thì có 5 Khoa và 1 Viện không cho Crawl Data về.



Hình 2.2 “robots.txt” của Facebook

1. Còn Crawler của Facebook chỉ cho phép các phần mềm Crawler uy tín như Googlebot, Bingbot,.. nên một điểm cần phải tìm hiểu thêm.

Cảm nhận: Khá là bổ ích về những kinh nghiệm và các cách sử dụng Crawler. Quan trọng biết được ”robots.txt” sẽ ảnh hưởng rất nhiều đến việc nghiên cứu. Và dữ liệu khi cào về được Import vào Database không nói tới.

Tóm tắt thông tin:

- Crawler là phần mềm có khả năng tự động lấy dữ liệu

- Có chức năng lấy thống tin từ Server, trích xuất ra thông tin người sử dụng cần, đồng thời tìm ra những link có trong trang Web đó và tự động truy cập vào những link đó

- Có nhiều tên gọi như robot, bot, spider, worm, ant

- Crawler tìm kiếm tự động của Google hay Yahoo kiểu đóng vai trò tìm kiếm link các trang Web Tiếng Việt

- Về mặt lý thuyết Crawler có vẻ khá đơn giản nhưng thực tế khi lập trình sẽ gặp phải khá nhiều vấn đề khó khăn và phức tạp

- Các bước xây dựng mô hình Crawler đơn giản:

+ Chọn URL khởi đầu

+ Sử dụng HTML Protocol để lấy trang Web

+ Trích xuất ra các Link. Lưu lại trong Queue

+ Lặp đi lặp lại bước 2,3

- Các vấn đề cần giải quyết:

+ Thời gian giới hạn: Nếu Server không trả lời thì chương trình sẽ bị đóng băng. Vì thế cần xử lý trường hợp Server không trả lời sau 1 khoảng thời gian quy định

+ Nếu không lên kế hoạch truy cập 1 cách hợp lý, chương trình sẽ liên tục truy xuất 1 trang Web, điều đó sẽ thành tấn công Dos và gây ra rất nhiều phiền toái vì thế phải quản lý tần suất truy cập 1 trang, ví dụ 1 request/1s hoặc ít hơn

+ Truy cập lại trang Web đã được xử lý xong. Nếu không xử lý vấn đề này chương trình sẽ bị rơi vào vòng lặp vĩnh viễn. Vì thế phải xây dựng phương pháp đánh dấu những link đã xử lý. Đơn giản nhất là lưu lại URL của những trang Web đã xử lý, trước khi thêm vào quêu 1 URL mới thì so sánh với những URL đã xử lý trước. Ngoài ra, có 1 cách làm nhanh hơn là sử dụng hash.

+ Các link khác nhau nhưng cũng chỉ 1 trang Web : Không thể sử dụng cách so sánh các link được. Vì thế phải viết thêm 1 hàm để biến đổi những link này về 1 dạng thống nhất (Normalize)

- Viêc sử dụng Crawler truy cập vào các Website thường gây nên tình trạng nghẽn mạch, vì thế các Website thường có quy định dành riêng cho các Crawler. Các quy định này thường được lưu vào file “robots.txt”

<https://topdev.vn/blog/crawling-du-lieu-tu-website-tim-hieu-ve-scrapingweb/>

Những điều quan trọng rút ra được:

1. Web Scraping và Web Crawling khác nhau ở chỗ một cái thu thập toàn bộ thông tin trang Website đó và một cái thu thập thông tin cần thiết với mục đích người dùng
2. Những điều cần lưu ý khi dùng Scraping Data:

Website đó sử dụng cơ chế chặn bot thông qua Request, IPs, Cookie, Captcha và các phần mềm thứ ba

Ghi dữ liệu quá nhiều ảnh hưởng đến Database của trang Website đó

Cấu trúc Website thay đổi thì cũng ảnh hưởng đến việc Scraping Data

Cảm nhận: hiểu hơn về mặt lý thuyết Web Crawling và Web Scraping, thấy được ở góc độ Admin của Trang Website đó không muốn bị đánh cắp dữ liệu thì có những cách nào, ở góc độ người lấy dữ liệu thì có những giải pháp nào để vượt qua những hàng rào bảo vệ dữ liệu. Và dữ liệu khi cào về Import vào Database có nói tới ở MySQL(Load Data Local Infile), MongoDB(mongoimport) các DB engines cũng hỗ trợ 2 dạng trên

Tóm tắt thông tin:

- Dữ liệu đều có đặc điểm chung là cần xây dựng mô hình học máy hoặc tìm kiếm dữ liệu mà không thể nhìn thấy bằng mắt thường.

- Khi bạn thu thập dữ liệu từ các nguồn sẵn có - báo hoặc các trang Website đến các cơ sở dữ liệu, sách nội bộ nếu thua thập bằng tay rất khó khăn vì 2 lý do: tỷ lệ sai lệch dữ liệu lớn và chậm chạp kèm theo việc tăng số lượng người xử lý dữ liệu sau khi lấy

- Thu thập thông tin: là việc xử lý các tập dữ liệu lớn nới cần các trình thu thập dữ liệu (hoặc bot) của riêng mình để thu thập các dữ liệu sâu nhất trên Website. Mặt khác, dữ liệu được dùng để lấy thông tin từ bất kỳ nguồn nào đều cần trích xuất dữ liệu dưới dạng Scraping (hoặc Harvesting) - Quan điểm sai lầm nghiêm trọng

- Scraping data

+ Scraping dữ liệu không nhất thiết phải liên quan đến Website

+ Scraping có thể đề cập đến việc trích xuất thông tin từ một hệ thống cục bộ, cơ sở dữ liệu chung hoặc thậm chí từ Internet

+ Web Scraping cũng thực hiện việc tìm kiếm và thu thập thông tin nhưng khác với Web Crawling, Web Scraping không thu thập toàn bộ thông tin của một trang Web mà chỉ thu thập những thông tin cần thiết, phù hợp với mục đích của người dùng

+ Trong Web Scraping chúng ta cũng phần nào sử dụng Web Crawler để thu thập dữ liệu, kết hợp với Data Extraction (trích xuất dữ liệu) để tập trung vào các nội dung cần thiết

|  |  |
| --- | --- |
| Data Scraping | Data Crawling |
| Involes extracting data form various sources including web | Refers to dowloading pages from the web |
| Can be done at any scale | Mostly done at a large scale |
| Deduplication is not necessarily a part | Deduplication is an essential part |
| Needs crawl agent and parser | Needs only crawl agent |

- Một số các trường hợp gặp phải khi Scraping Data:

+ Anti-bot detection : Khi cào thì gặp Website sử dụng cơ chế chặn bot( nhưng không chặn các Search Engine)

1. Phát hiện User-Agents truy cập nhiều: giải pháp dùng user-agent khác nhau hoặc dùng SE Agents nổi tiếng cho mỗi lần request. Danh sách 450 User-Agents dowload tại : [http://cafemmo.club/threads/chia-se-danh-sach-user-agent-thong-dung-nhat.1818/](http://cafemmo.club/threads/chia-se-danh-sach-user-agent-thong-dung-nhat.1818/" \t "https://topdev.vn/blog/crawling-du-lieu-tu-website-tim-hieu-ve-scrapingweb/_new)
2. Phát hiện IPs truy cập nhiều, giả sử 5 requests/s: giải pháp dùng các dịch vụ IP rotator cho mỗi lần request. Ở đây dùng stormproxies.com
3. Phát hiện người dùng thật qua JavaScript, đa số bot không hỗ trợ JS mà: giải pháp dùng headless browser như Splash, Selenium, PhantomJS, Puppeter,.. Có khá nhiều sites dùng JS để detect robot
4. Sử dụng honeypot traps: ví dụ như các links bẩy đính kèm display: none, visibility: hidden,.. => cài đặt cơ chế phát hiện các traps
5. Sử dụng cookie, captcha để chặn, đa số sites dùng Cloudflare để chặn bot: có vài Script Bypass Cloudflare rồi sử dụng [https://github.com/Anorov/cloudflare-scrape](https://github.com/Anorov/cloudflare-scrape" \t "https://topdev.vn/blog/crawling-du-lieu-tu-website-tim-hieu-ve-scrapingweb/_self) (Script này by pass cookie của Cloudflare nhưng chưa có cơ chế bypass captcha của Cloudflare)

+ Ghi kết quả dữ liệu (Scraped data) quá nhiều

Khi bạn cào dữ liệu với nhiều Spiders thì việc ghi dữ liệu quá nhiều và liên tục vào DB sẽ làm cho DB của bạn quá tải và giảm hiệu năng có thể ảnh hưởng đến hiệu suất hoạt động. Nên cân nhắc:

1. Bulk insert query: tức sipder chỉ cần thực hiện 1 query để insert nhiều records
2. Bulk import file: tức là spider ghi dữ liệu vào 1 file với 1000 dữ liệu chẳng hạn, sau đó sử dụng lệnh import file đó vào DB

+ Cấu trúc site thay đổi: như thay đổi Layout tức HTML tags thay đổi, lúc này bạn phải thay đổi các Selectors để lấy đúng dữ liệu bạn cần. Vì thế các Spider cần có cơ thế phát hiện sự thay đổi cấu trúc site để thông báo và dừng extract data của site đó. Khi đó, mỗi site ta cần hỗ trợ nhiều Schemas để Extract dữ liệu hơn

- Cách ngăn chặn site Scraping

+ Sử dụng công cụ phân tích

+ Triển khai các challenge-based để đánh giá hành vị của người dùng nếu hỗ trợ cookie và javascipt

+ Lựa chọn hành vi tiếp cận dữ liệu

+ Sử dụng robots.txt để bảo vệ Website trước Scraping bot(hướng dẫn các con bot thực hiện theo rules định sẵn). Tuy nhiên phương pháp này không được lâu dài

Nguồn: [https://securitydaily.net/tim-hieu-ve-web-scraping-bot-la-gi/](https://securitydaily.net/tim-hieu-ve-web-scraping-bot-la-gi/" \t "https://topdev.vn/blog/crawling-du-lieu-tu-website-tim-hieu-ve-scrapingweb/_new)[https://www.promptcloud.com/blog/data-scraping-vs-data-crawling/](https://www.promptcloud.com/blog/data-scraping-vs-data-crawling/" \t "https://topdev.vn/blog/crawling-du-lieu-tu-website-tim-hieu-ve-scrapingweb/_new)

<https://lptech.asia/dich-vu/dich-vu-lap-trinh-cong-cu-crawl-du-lieu-website-tu-dong>

Những điều quan trọng rút ra được:

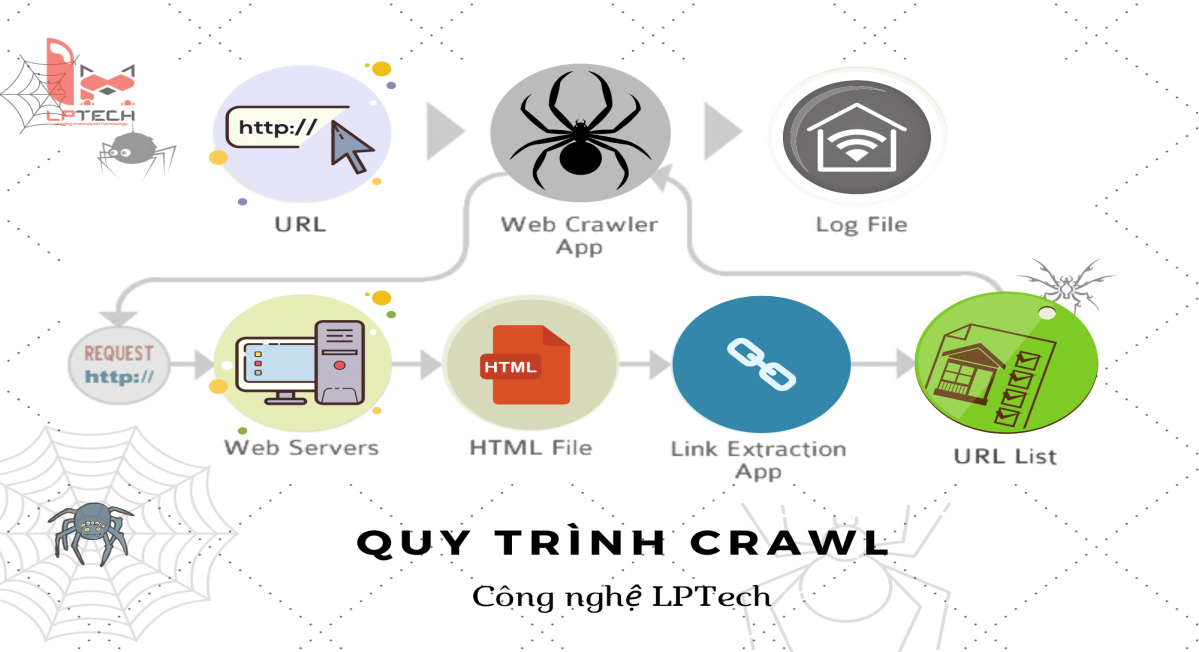
1. Việc Crawl Data từ một trang Webite có quyền sở hữu trí tuệ, quyền tác giả và quyền liên quan mà chưa được sự chủ quyền cho phép thì có thể vi phạm Luật Pháp Việt Nam
2. Phần mềm Crawl Data có rất nhiều lợi ích đối với các doanh nghiệp cần nhiều thông tin, nhiều nội dung ví dụ như Thương Mại Điện Tử, Trang tin tức
3. Có 3 ngôn ngữ lập trình đối với Crawler là tốt nhất: Python, Node, PHP. Và một công cụ AI để xử lý phân tích các Website cao cấp thay đổi cấu trúc trang liên tục và xử lý nội dung trước khi Crawl về để tránh trùng lắp nội dung 100%.

Cảm nhận: Crawl Data không chỉ là việc cào dữ liệu về. Mà nó còn là một lá bài để các doanh nghiệp có được những dữ liệu quan trọng và ai biết nó còn là một trận chiến ai sẽ trở thành Content IS King. Và dữ liệu khi cào về được Import vào Database không được nói tới.

Tóm tắt nội dung:

- Crawl dữ liệu hay còn gọi là cào dữ liệu là một thuật ngữ không có gì xa lạ trong ngành marketing

- Crawler có công việc chính là thu thập dữ liệu từ một trang Website bất kì, hoặc chỉ định trước rồi phân tích cú pháp mã nguồn HTML để đọc dữ liệu và bóc tách thông tin dữ liệu theo yêu cầu mà người dùng đặt ra hoặc các dữ liệu mà Search Engine yêu cầu.



Thời đại hiện tại là thời đại của công nghệ 4.0 và Big-Data cho nên bạn càng làm chủ Data bao nhiêu thì bạn càng có nhiều cơ hội trong việc thương mại của doanh nghiệp của bạn hoặc cho cá nhân bạn vào một ứng dụng nào đó.

Crawler phù hợp với những doanh nghiệp nào

- Sàn TMĐT, Website rao vặt.

- Tin tức hàng / hằng ngày

- Pháp luật đời sống

- Website vệ tinh - PBN

- Website bán hàng Online, Nhập hàng nước ngoài

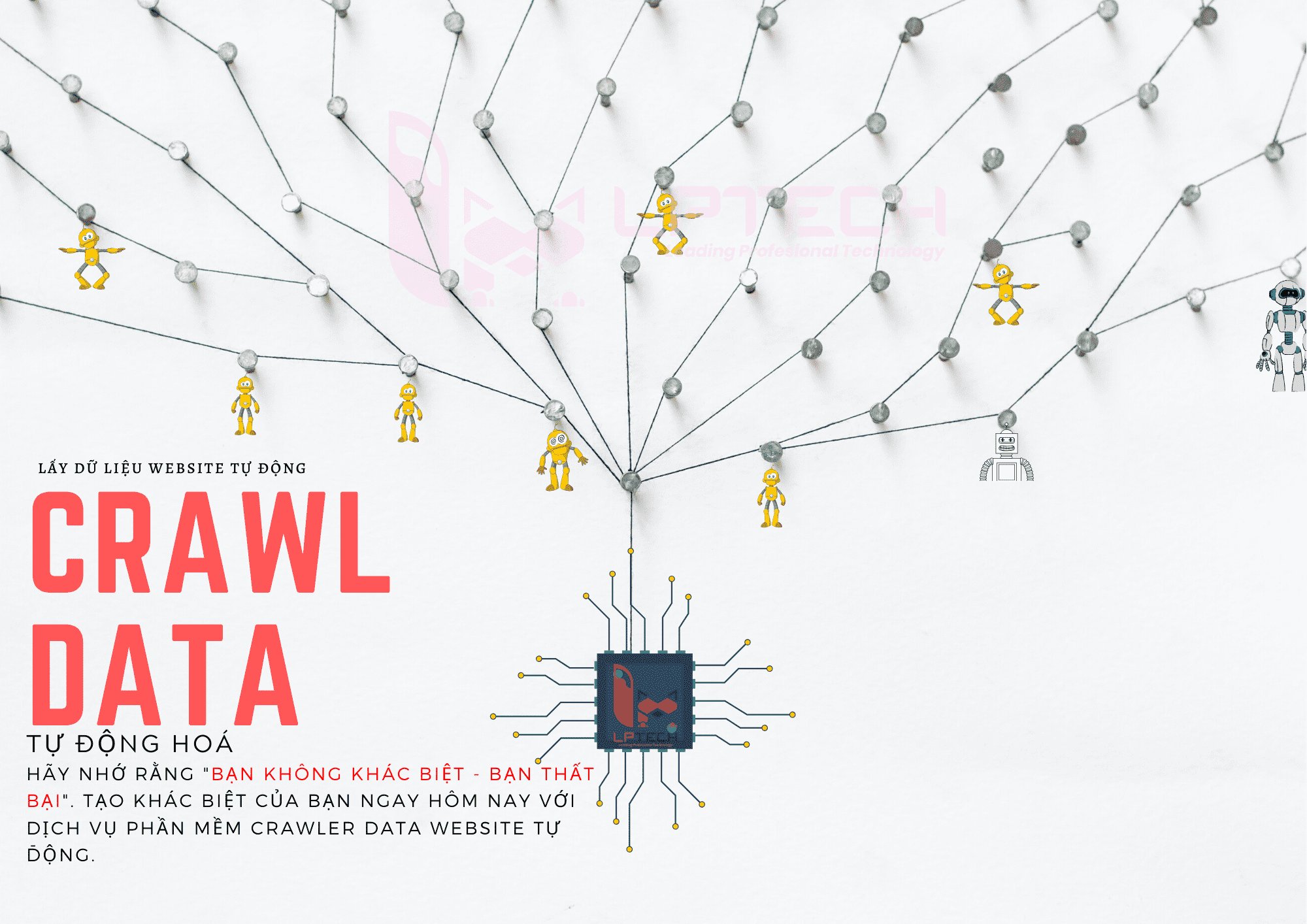
Việc phát triển 1 công cụ Crawl dữ liệu cũng tốn một ích chi phí, cho nên việc này cũng cần công ty của bạn có năng lực tài chính ổn, xem thêm ở phần chi phí

Lợi ích của việc Crawler Data

- Giảm tải công việc sáng tạo cho nhân viên Content của bạn

- Làm giàu nội dung, nhiều tin tức của Trang Website thay bằng phải thuê 1 đội nhập liệu vài trăm nhân sự sẽ rất tốn chi phí, cồng kềnh và thủ tục pháp lý đi kèm cho nhân sự không hề đơn giản. Sẽ có nhiều Users hơn

- Đầu tư một phần mềm Crawler data tự động thì bạn có thể giảm tải gần như 90% nhân sự content hiện tại, chỉ giữ 10% nhân sự để chỉnh sủa, viết lách các nội dung quan trọng cho công ty và quản trị công cụ Crawler data mà thôi



Bật mí bí mật

Các công ty chuyên bán hàng bằng Affiliate (Tiếp thị liên kết) thì việc cần 1 công cụ Crawl Link, Crawl Data là vô cùng quan trọng, bạn chỉ cần Crawler hết Data của các sản phẩm ở Website khác, sau đó gắn Link ?Ref=Code(Refer) để có thể tăng doanh số của mình 1 cách chóng mặt

Công nghệ sử dụng là gì?

Công cụ mới nhất hiện nay để Crawl và bóc tách dữ liệu một cách chính xác và thông tinh. Các ngôn ngữ lập trình Crawler tốt nhất hiện nay:

Python

PHP

Node

Proxy trong Crawl là điều vô cùng quan trọng chống các Website Victim chặn việc thu thập của chúng ta, ngoài ra còn có các kỹ thuật khác sử dụng AI để phân tích Website cao cấp và có cấu trúc thay đổi liên tục như Zalo Shop, Tiki, Sendo, ChoTot, MuaBan,..

Có bị phạt không?

Đây là một vấn đề gặp phải của các công ty phần mềm cung cấp dịch vụ này

Đối với Google: Việc copy hay Crawl tạo ra 1 bản sao chép Website đó về Database của bạn nếu bạn chỉ Crawler 100% nội dung thì có thể bạn vi phạm chính sách nội dung của Google và DMCA sẽ khởi kiện bạn. Để giải quyết thì phải có một công cụ thông minh để Xử Lý dữ liệu 1 lần trước khi Crawl về nhằm tránh trùm lắp nội dung. Lưu ý hay không ngoan vì thuật toán của Google phát hiện chặn sớm. Giải quyết là AI

Đối với pháp luật Việt Nam: Việt Nam có luật bản quyền tác giả được công bố tại Nghị Định 22/2018/NĐ-CP quy định chi tiết Luật Sở Hữu trí tuệ, Luật sửa đổi Luật Sở hữu trí tuệ về quyền tác giả, quyền liên quan.

Giải quyết là phải được đơn vị chủ quyền cho phép, phải thay đổi nội dung gốc để tránh vi phạm

Kết luận Việc Crawl dữ liệu để Website, ứng dụng của bạn có càng nhiều thông tin hữu ích càng quan trọng vì việc đó càng tăng khả năng tiếp cận của bạn đến mọi người trên Internet. Tất cả các doanh nghiệp lớn làm TMĐT, rao vặt như MuaBan, ChoTot, muabanhanh… đều có hệ thống Crawl Data của họ. Vì càng có nhiều Content thì bạn càng làm chủ vì ConTent IS King

<https://toidicodedao.com/2015/07/28/tutorial-trich-xuat-thong-tin-tu-website-voi-html-aglitity-pack/>

Những điều quan trọng rút ra được:

1. Nếu sử dụng C# để lấy dữ liệu từ một trang Website thì sử dụng 2 bộ thư viện là HTML AgilityPack và Fizzler.
2. Xem xét HTML của trang Website cần trích xuất. Nên chọn Tag dựa theo id

Cảm nhận: thấy được Web Crawler sử dụng C# thì nên làm như thế nào và người viết cũng nói rất chi tiết kể cả những lỗi sẽ gặp phải. Và dữ liệu khi cào về Import vào Database có nói tới ở B6 của người viết.

Tóm tắt nội dung:

Trích xuất thông tin từ Website với HTML Aglitity Pack

Hiện nay nhu cầu thu thập dữ liệu ngày càng tăng. Với một số trang lớn như Facebook, Google, Steam ta có thể sử dụng API do họ cũng cấp để lấy dữ liệu. Trong nhiều trường hợp khác, ta thường trích xuất dữ liệu bằng tay (Word, Excel) vừa cực vừa mất nhiều thời gian

Không có API để lấy dữ liệu. Ở đây, ta không thể lấy dữ liệu bằng tay được vì mất rất nhiều thời gian và dẫn đến sai sót. Giải pháp duy nhất cho chuyện này là viết một phần mềm trích xuất dữ liêu từ bản thân

Sẽ dùng 2 bộ thư viện:

+ HTML AgilityPack: là một thư viện parse HTML khá mạnh vì chơi được với hầu hết HTML, cả vaild và unvalid. Trong thực tế thì số lượng Website có HTML unvalid nhiều vô số kể, các thư viện khác sẽ dễ bị lỗi.

+ Fizzler

+ Từ khóa: Web Crawler

B1: Tạo Project Console đơn giản

B2: Cài đặt thư viện 2 thư viện trên

B3: Xem xét HTML của trang cần trích xuất. Kinh nghiệm của mình là bạn nên chọn tag dựa theo id, nằm gần dữ liệu mình cần lấy nhất. Vì mỗi tag trong html chỉ có 1 id duy nhất không bị trùng ta dễ chọn tag và lọc hơn

B4: Giới thiệu 1 số object, method của HTML Agility Pack

+ HTML Document: đây là một class chứa thông tin về một file HTML (encoding, innerhtml). Ta có thể load dữ liệu vào HTML Document từ 1 URL hoặc từ 1 file.

+ HTML Node: một HTMLNode tương đương với một tag(li,ul,div,..) trong HTML. Node lớn nhất chứa toàn bộ tất cả sẽ là DocumentNode. Một số property của HTMLNode mà hay sử dụng

Name: tên của Node(div,ul,li)

Attributes: danh sách các attribute của note(attribute là thông tin của node như src,href,id,class)

InnerHTML, OuterHTML

SelectNodes(string xPath): tìm các node con của node hiện hành, dựa trên xPath đưa vào

SelectSingleNode(string xPath): tìm node con đầu tiên của node hiện hafnhm dựa trên xPath đưa vào

Descendants(string xPath): trả ra danh sách các HTMLNode con của node hiện tại

Đầu tiên ta sử dụng method SelectNode, sử dụng xPath để tìm Node hoặc dùng LINQ to Object để tìm note.

Nếu dùng LINQ to Object thì phải dùng Lambda Expression và check null trước để tránh bị lỗi NullPointerException. Nếu bạn bỏ vế check null ra sau, hàm chạy sẽ bị lỗi

B5: Cải tiến với Fizzler

- Có một cách đơn giản để Select 1 Node, đó là sử dụng Fizzler. Fizzler hỗ trợ CSS Selector, cho phép ta sử dụng Selector của CSS. Fizzler được mở rộng dựa trên HTMLAglityPath, thêm 2 hàm sau vào HTMLNode:

+ QuerySelectorAll: Tìm các node con của node hiện hành, dựa trên CSS Selector đưa vào

+ QuerySelector: Tìm node con đầu tiên của node hiện hành, dựa trên CSS Selector đưa vào

B6: Xuất kết quả ra đâu đó. Tới đây mọi chuyện đã xong, có thể lưu kết qủa vào file database hoặc xuất ra file text tùy mục đích sử dụng

<https://codetheworld.io/nodejs-dung-selenium-cheerio-de-crawl-du-lieu.html>

Những điều quan trọng rút ra được:

1. Những trang Website chạy bằng Ajax hay phải qua bước Login thì sẽ khó khăn khi phải Crawl dữ liệu từ trang Website đó về. Và Selenium là lựa chọn thích hợp nhất để khắc phục chuyện đó
2. Thường hay bị lỗi ở Requyest Timeout phải Restart Container. Một điều cần phải lưu ý

Cảm nhận: Nội dung bài viết khá là hay sử dụng công nghệ Selenium để Crawl dữ liệu về. Và dữ liệu khi cào về được Import vào Database có nói tới là lưu dữ liệu vào file json và nạp vào Database.

Tóm tắt nội dung

Một cách khá đơn giản đó là phân tích HTML, đọc các thẻ và rút trích dữ liệu

Việc Crawl một trang bằng đọc HTML thuần sẽ rất khó khăn trong một số trường hợp như: dữ liệu được load bằng Ajax (lúc đọc HTML sẽ chỉ thấy Wrapper chứ không thấy dữ liệu, dữ liệu Ajax trả về bị Encode) hay muốn vào được trang cần Crawl thì phải qua bước Login

Những trang như vậy khi chạy Javacript sẽ xử lý đổ dữ liệu và sau đó mới đổ vào cây DOM. Khi dùng Request đọc HTML thì sẽ không thấy được các thẻ DIV như khi Inspect Element

Với những loại như vậy mình sử dụng Selenium để chạy Website trên Browser thật, thực hiện thao tác để được trang HTML fully load rồi mới trích xuất dữ liệu.

Selenium khá nổi tiếng trong lĩnh vực Automation Test. Nó cho phép mình chạy Script Test trên Browser thật. Cách làm của mình sẽ là: dùng Selenium chạy trang, đợi Javascript chạy xong và sau đó Crawl dữ liệu bình thường

Vì Selenium chạy trên nền JVM(máy tính cần cài đặt môi trường Java), nên mình tìm trên mạng giải pháp bằng docker: Selenium Docker

Selenium Grid

- Hệ thống này chuyên dùng cho việc thực hiện test tự động: thực hiện nhiều test case, trên nhiều trình duyệt, version trình duyệt khác nhau, trình duyệt trên các hệ điều hành khác nhau.

- Hệ thống Selenium Grid bao gồm một Hub và một hoặc nhiều node(không phải nodejs). hub là trung tâm tiếp nhận các yêu cầu thực hiện test, thông tin này bao gồm test case, loại trình duyệt, hệ điều hành nơi các test sẽ được chạy. Các node sẽ kết nối vào Hub nhận các Request từ Hub. Như vậy node sẽ là nơi chạy các trình duyệt và thực hiện các test case

Sử dụng phương pháp này chúng ta có thể giải quyết nhiều bài toán crawl dữ liệu

Trong quá trình chạy thấy tốc độ hơi chậm và thỉnh thoảng WebDriver bị Request Timeout, phải Restart các Container

PHÂN TÍCH TRANG WEBSITE HUTECH TUẦN 2

Trang chủ Hutech

<https://www.hutech.edu.vn/homepage/index.php/tin-tuc/tin-hutech>

<https://www.hutech.edu.vn/homepage/index.php/tin-tuc/hoat-dong-sinh-vien>

Phân tích những bài biết thì có thể thấy

- Những tiêu đề, ảnh sự kiện được lặp đi lặp lại, phân tích cấu trúc HTML thì với class=”news-title” và ảnh class=”img-responsive” được sử dụng lại nhiều lần nên việc cào về sẽ không có gì khó khăn nhưng vẫn thiếu những thông tin dữ liệu cần thiết như thời gian chẳng hạn

<https://www.hutech.edu.vn/homepage/tin-tuc/hoat-dong-sinh-vien/14583525-gia-han-dang-ky-cuoc-thi-sinh-vien-khoi-nghiep-hutech-startup-wings-2020-den-3103>

Phân tích bài viết trên có thể thấy

- Tiêu đề sự kiện: ở phần cấu trúc HTML được đặt class=”title-block”

- Ảnh sự kiện: ở phần cấu trúc HTML thì chỉ dùng style không có id hay class

- Còn lại các dữ liệu khác ở phần cấu trúc HTML không có id hay class

Việc tách bóc nội dung dữ liệu không có id hay class là một rủi ro vì nếu viết code tách bóc dành cho 1 bài viết đó thì những bài viết khác thì không giống vậy nên phải viết lại lần nữa gây mất thời gian và không hiệu quả

Nếu tách bóc chỉ tên sự kiện và ảnh sự kiện thì có thể phải thay đổi hệ thống ràng buộc