|  |  |  |  |
| --- | --- | --- | --- |
| **THÔNG TIN THÀNH VIÊN NHÓM** | | | |
| **#** | **Họ tên** | **MSSV** | **Lớp** |
| 1. | Lê Khôi Đệ | 1411010102 | 15DTH06 |
| 2. | Nguyễn Bá Tâm | 1611061022 | 16DTHC2 |
| 3. |  |  |  |

**NỘI DUNG BÁO CÁO TUẦN 2**

Báo cáo các nội dung bao gồm:

* Nội dung đã thực hiện: phân tích nghiên cứu Crawl Data
* Nội dung đang thực hiện: phân tích nghiên cứu khi cào dữ liệu về thì Import vào Database và phân tích trang Website Hutech
* Khó khăn: Khi phân tích HTML trang Website Hutech thì chỉ có tiêu đề sự kiện, ảnh sử kiện là dễ dàng lấy vì cấu trúc HTML lặp lại nhiều lần. Còn những dữ liệu khác(thời gian, cơ sở,…) thì phức tạp hơn phải cần thời gian nghiên cứu kĩ hơn
* Câu hỏi: Em không có câu hỏi gì cả.

**2.1 WBS (Cấu trúc phân rã công việc - cập nhật liên tục)**

*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

**2.2 Dữ liệu khi cào về được Import vào Database(Cập nhật liên tục)**

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

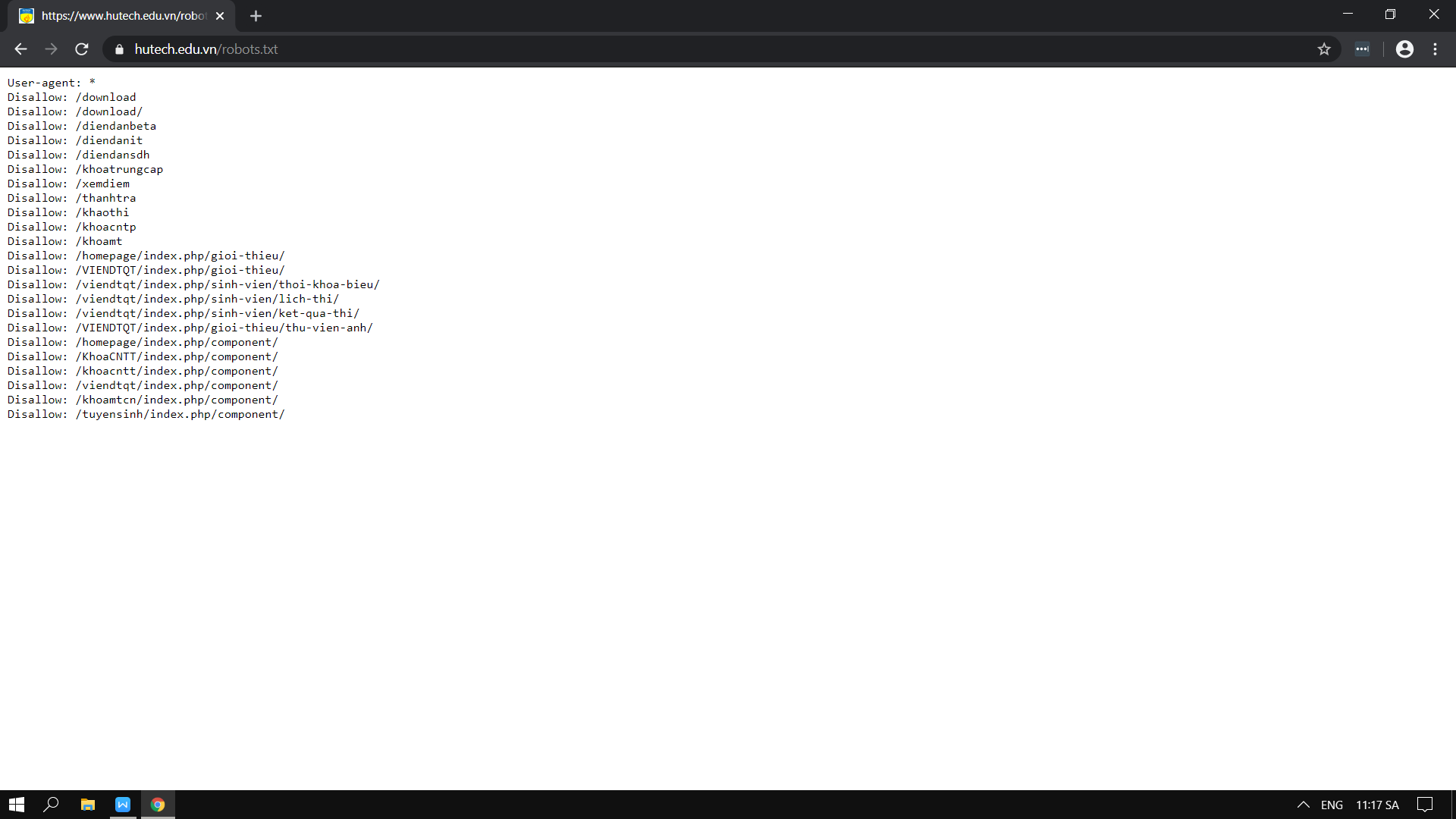
Điều quan trọng rút ra được:Trước khi Crawl một trang Web thì phải soi toàn bộ cấu trúc HTML để đưa ra phương án nghĩa là mỗi trang phải code khác nhau vì cấu trúc khác nhau

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

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

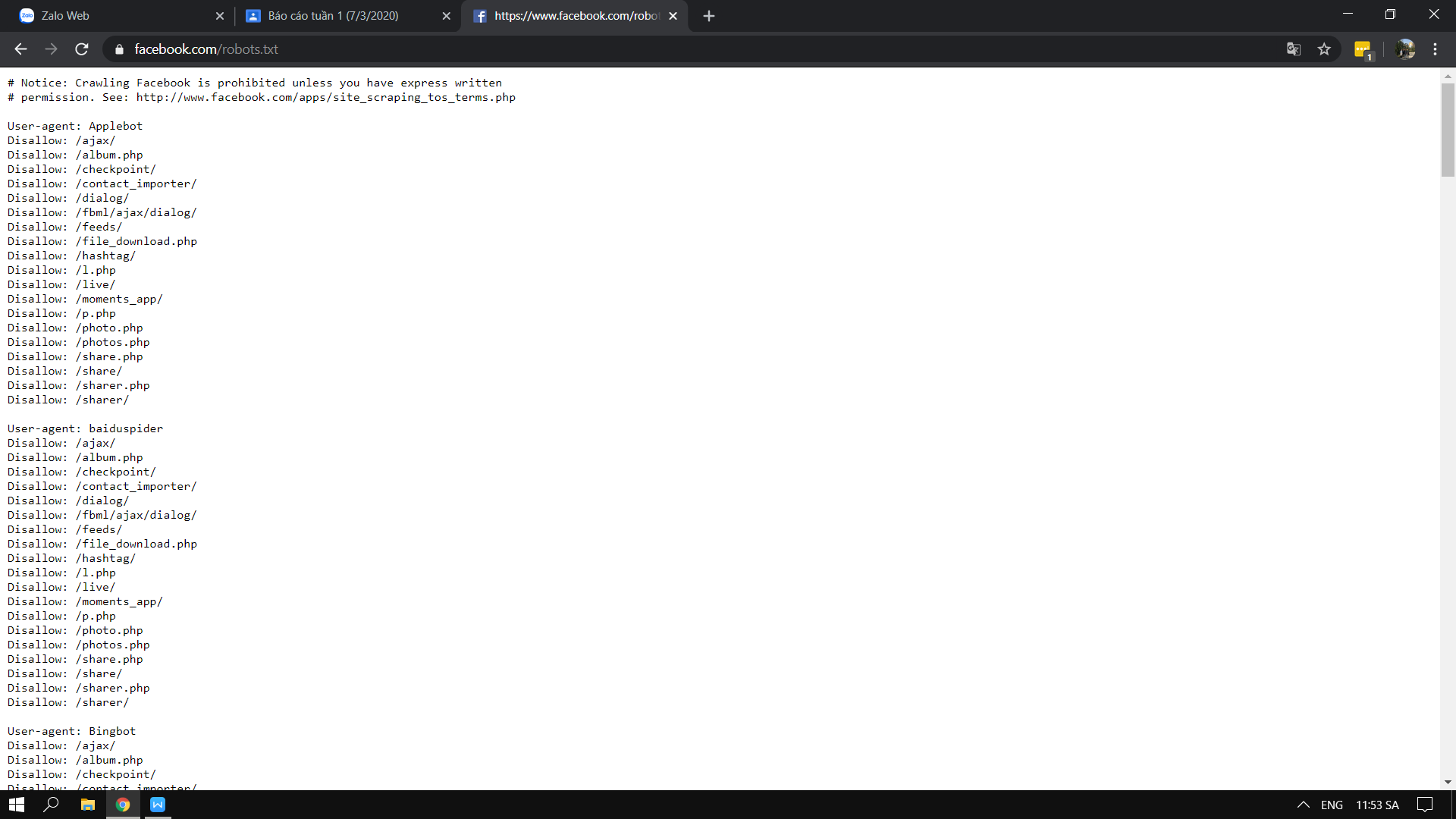
Những điều quan trọng rút ra được:

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.

<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

<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.

<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.

<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.

**2.3 Phân tích trang Webiste Hutech (Cập nhật liên tục)**

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

Kết luận: 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 có cấu trúc không giống như 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 để có thể tạo được sự kiện