

ĐẠI HỌC BÁCH KHOA HÀ NỘI
ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG



BÁO CÁO MÔN HỌC
IT3150 – Project 1

Thu Thập thông tin bài báo từ trang web điện tử

Giảng viên hướng dẫn: **PGS. TS Lê Thanh Hương**

Sinh viên thực hiện : Lương Triều Vỹ

MSSV : 20204708

Lớp : 721012

Hà Nội, tháng 03 năm 2023

Nội dung

I.	Giới thiệu đề tài.....	3
II.	Công nghệ sử dụng	3
III.	Các chức năng của ứng dụng.....	3
1.	Trích xuất thông tin từ các trang web tin tức 24h và express	3
2.	Lưu thông tin các bài báo vào file trên máy tính.....	4
3.	Tự động hóa việc thực hiện nhiệm vụ hàng ngày bằng ứng dụng Task Scheduler.....	4
IV.	Cách thức thực hiện	4
1.	Trang báo 24h.....	4
a.	Xử lý trang báo	4
b.	Xử lý bài viết	6
2.	Trang báo Express	7
a.	Xử lý trang báo	7
b.	Xử lý bài báo	8
3.	Sử dụng Task Scheduler	9
V.	Kết quả đạt được	11
VI.	Hướng phát triển tiếp theo:.....	11

Lời nói đầu

Báo chí từ lâu luôn là một kênh thông tin quan trọng phản ánh mọi vấn đề, sự việc trong xã hội. Và với sự phát triển của Internet và các thiết bị di động hiện nay, báo điện tử đã trở thành một loại phương tiện truyền thông phổ biến, cho phép người dùng có thể truy cập và đọc các tin tức, thông tin từ mọi nơi trên thế giới.

Sở hữu nhiều lợi thế so với báo giấy truyền thống như: dễ dàng truy cập, nội dung đa dạng, cập nhật nhanh chóng và tính tương tác cao, báo điện tử phát triển khá nhanh chóng và thay đổi thói quen đọc báo của không ít người đọc. Tuy nhiên, báo điện tử cũng tồn tại nhiều mặt trái khi những nội dung không lành mạnh, độc hại có thể được lan truyền nhanh chóng.

Một số trang báo điện tử lớn tại Việt Nam có thể kể đến như: Vnexpress, 24h, Dân trí, Zingnews, Báo thanh niên, v.v...

Việc thu thập dữ liệu từ các báo điện tử có tác dụng quan trọng trong việc phân tích, đánh giá và dự báo các xu hướng của thị trường, kinh tế và chính trị, xã hội. Cùng với đó cũng có thể được sử dụng để phát hiện và theo dõi các thông tin sai lệch, tin tức giả và các nội dung độc hại khác.

I. Giới thiệu đề tài

Trong đề tài này, chúng tôi đã xây dựng một ứng dụng trích xuất thông tin từ các trang web tin tức 24h và vnexpress bằng cách sử dụng ngôn ngữ lập trình Java và công cụ Jsoup. Đồng thời, chúng tôi đã tự động hóa việc thực hiện nhiệm vụ này hàng ngày bằng cách sử dụng ứng dụng Task Scheduler

II. Công nghệ sử dụng

- Ngôn ngữ lập trình : Java
- Công cụ trích xuất thông tin: Jsoup
- Môi trường: Eclipse
- ứng dụngTask Scheduler: Windows Task Scheduler
- Công cụ truy xuất : cssSelector

III. Các chức năng của ứng dụng

1. Trích xuất thông tin từ các trang web tin tức 24h và express

- Lấy danh sách các bài báo từ trang chủ của 24h và express
- Truy cập vào từng bài báo và lấy thông tin: tiêu đề, ngày đăng, id , tác giả và nội dung bài báo

2. Lưu thông tin các bài báo vào file trên máy tính

- Tạo thư mục để lưu trữ các bài báo được trích xuất
- Lưu thông tin của từng bài báo vào các file với định dạng .json

3. Tự động hóa việc thực hiện nhiệm vụ hàng ngày bằng ứng dụng Task Scheduler

- Thiết lập kế hoạch thực hiện nhiệm vụ hàng ngày trong Windows Task Scheduler
- Ứng dụng sẽ tự động chạy vào mỗi ngày lúc cố định và thực hiện các bước trích xuất thông tin và lưu vào file

IV. Cách thức thực hiện

1. Trang báo 24h

a. Xử lý trang báo

Đối với trang báo 24h, ta cần làm quen với 1 khái niệm là robots.txt.

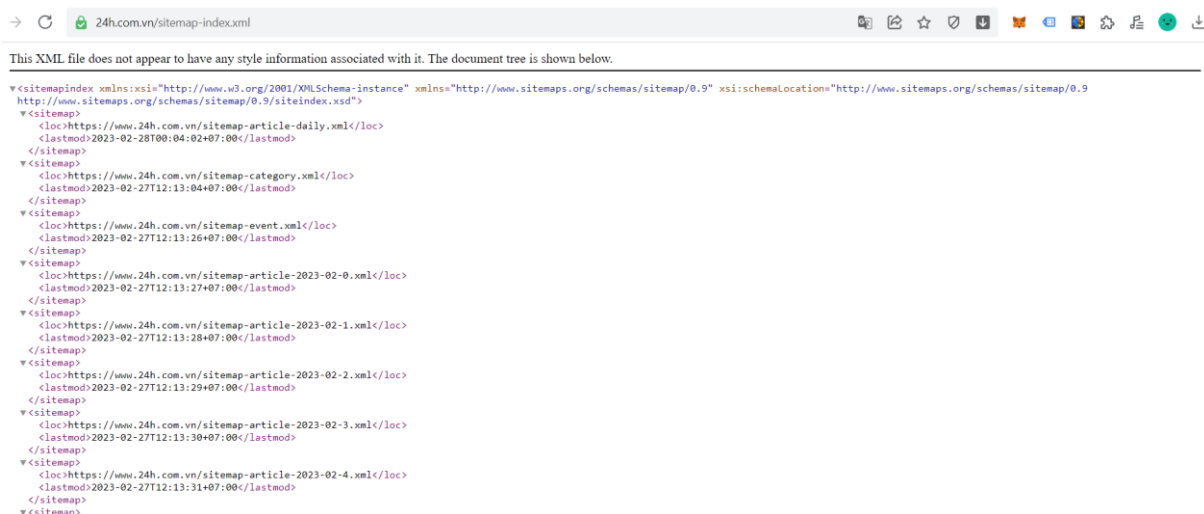
Robots.txt là một tập tin văn bản thuần túy, có định dạng **.txt** và nằm trong thư mục gốc của website. File robots.txt này quy định về việc truy cập, thu thập dữ liệu và index nội dung trên website của bạn đối với các công cụ tìm kiếm như Google, Bing, Cốc Cốc,...

File robots.txt quyết định về việc thu thập hoặc không thu thập dữ liệu của các thành phần có trên website của bạn. Nhờ đó, kiểm soát và tối ưu việc thu thập dữ liệu, giúp website của bạn không bị đánh giá xấu trước các công cụ tìm kiếm. Đồng thời, giữ nội dung/ thành phần trên website của bạn riêng tư ở một mức độ nào đó.

```
#User-agent: *
#Disallow: /

User-agent: *
Allow: /
Disallow: /ocm/
Disallow: /ad/
Disallow: /tools/
Disallow: /webservices/
Disallow: /crondamon/
Disallow: /trienkhai/
Disallow: /124557882/
Disallow: /su-kien/
Disallow: /*recommend-video_news
Sitemap: https://www.24h.com.vn/sitemap-index.xml
```

Sơ đồ trang web (sitemaps) là danh sách các bài viết, trang hoặc những tập tin có trên Website. Chúng được sắp xếp theo thứ tự theo sơ đồ phân tầng theo từng danh mục, theo thời gian đăng bài hoặc thời gian chỉnh sửa bài viết.



Khi truy nhập vào đường link: <https://www.24h.com.vn/sitemap-index.xml> , ta có thể thấy thể

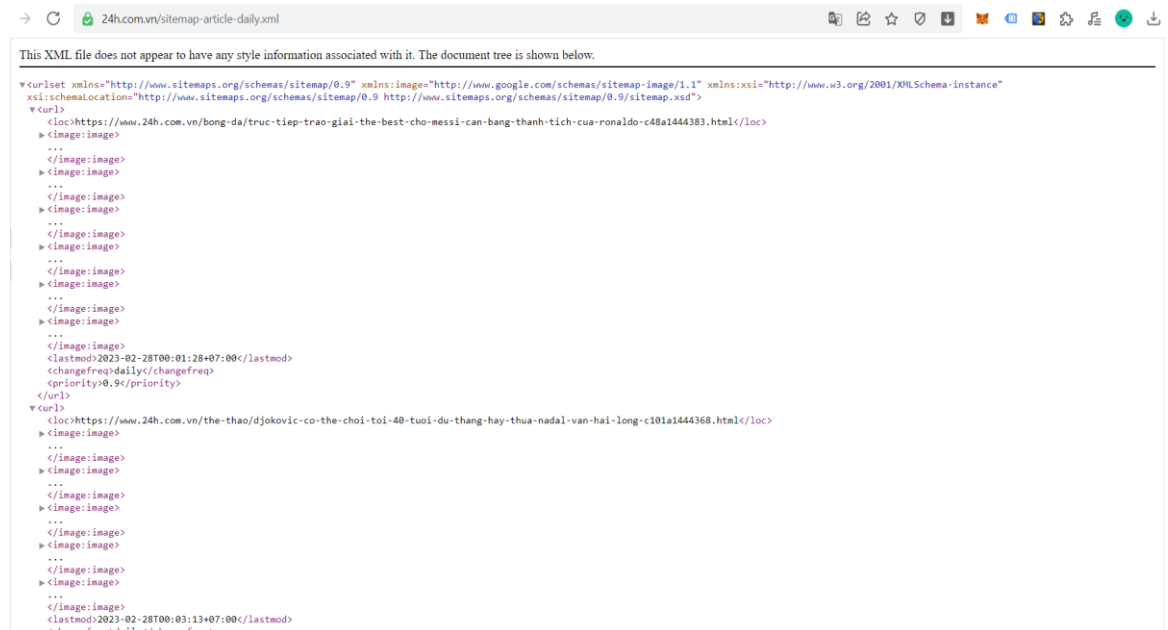
<sitemap>

<loc>https://www.24h.com.vn/sitemap-article-daily.xml</loc>

<lastmod>2023-02-28T00:04:02+07:00</lastmod>

</sitemap>

Đó là đường link dẫn tới danh sách bài viết của ngày hiện tại, khi truy nhập tới đường link <https://www.24h.com.vn/sitemap-article-daily.xml>



Ta có thể truy xuất được 2 dữ liệu là :

- Đường link bài báo thông qua thẻ <loc>
- Thời gian của bài viết thông qua thẻ <lastmod>

b. Xử lý bài viết

Sau khi có đường link dưới dạng text của bài viết, ta sử dụng câu lệnh sau:

```
Document doc_bai = Jsoup.connect(link).get();
```

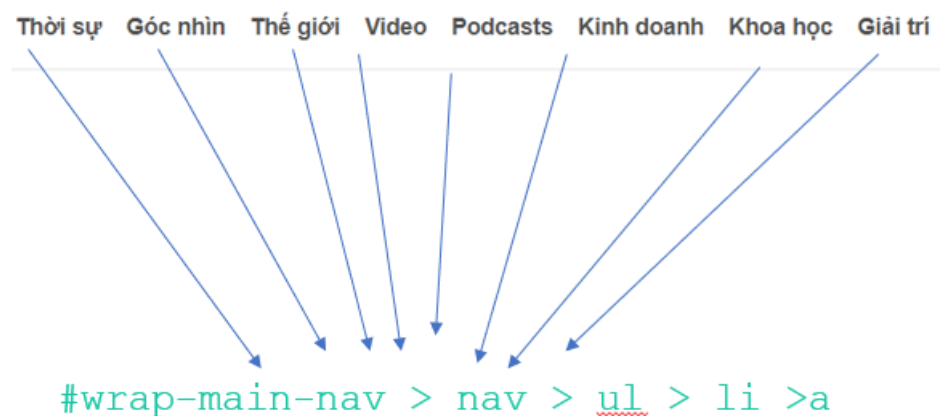
Thông tin	Cách thức truy xuất
Id	Đường link bài báo
Ngày viết	Thẻ <lastmod>

Tác giả	Class = <code>nguontin nguontinD bld mrT10 mrB10 fr</code>
Tiêu đề	Phương thức <code>titile</code>
Nội dung	Thẻ <code>#article_body > article.cate-24h-foot-arti-deta-info > p</code>

2. Trang báo Express

a. Xử lý trang báo

Đối với trang báo express, ta thu thập dữ liệu theo chiến lược bắt đầu từ trang chủ, sử dụng `cssSelector` để lấy các thư mục con của trang báo



Từ mỗi thư mục con, ta thu nhập hết tất cả các đường link của trang thư mục con đó và phân loại thành 3 kiểu đường link

1, đường link thư mục

Trong các trang thư mục con bao gồm: đường link bài báo, đường link thư mục tiếp theo, đường link ngoại lệ

Vd: <https://vnexpress.net/thoi-su-p2>, <https://vnexpress.net/thoi-su-p3> ,...

2, đường link bài báo

Đường link bài báo thường có id là duy nhất ở cuối mỗi đường link, ta có thể trích xuất id từ ngay chính đường link có tác dụng một bài báo được thu thập dữ liệu nhiều lần

Vd: <https://vnexpress.net/khoi-cong-duong-lien-ket-vung-hoa-binh-ha-noi-cai-toc-son-la-4574995.html> sẽ có id là 4574995

3, đường link ngoài lề

Đường link ngoài lề là những đường link để dẫn đến các trang báo khác hoặc dẫn đến bình luận của bài báo...

Vd: https://vnexpress.net/khoi-cong-duong-lien-ket-vung-hoa-binh-ha-noi-cai-toc-son-la-4574995.html#box_comment_vne

b. Xử lý bài báo

Trước khi trích xuất nội dung bài báo, ta cần đảm bảo:

- Bài báo chưa được thu nhập dữ liệu
- Bài báo được viết trong ngày hiện tại

Thông tin	Cách thức truy xuất
Id	Đường link bài báo
Ngày viết	Thẻ <code>#dark_theme</code> > <code>section.section.page-detail.top-detail</code> > <code>div</code> > <code>div.sidebar-1</code> > <code>div.header-</code> <code>content.width_common</code> > <code>span</code>
Tác giả	Thẻ <code>#dark_theme</code> > <code>section.section.page-detail.top-detail</code> > <code>div</code> > <code>div.sidebar-1</code> > <code>article</code> > <code>p.author_mail</code> > <code>strong</code>


Tiêu đề	Phương thức title
Nội dung	<div>Thẻ <code>#dark_theme</code> ></div> <div><code>section.section.page-detail.top-detail</code> ></div> <div><code>div</code> > <code>div.sidebar-1</code> > <code>article</code> ></div> <div><code>p.Normal</code></div>

3. Sử dụng Task Scheduler

Các bước thực hiện:

Bước 1: xuất chương trình java thành file .jar

 `crawl_24h_Main.jar`

 `crawl_express_Main.jar`

Bước 2: tạo file .cmd và .bat

Nội dung file cmd: `java -jar` + đường dẫn file jar

Nội dung file bat: đường dẫn file cmd tương ứng

```
run_cmd_24h.cmd - Notepad
File Edit View
java -jar "C:\Users\vykro\Desktop\all-jar\crawl_24h_Main.jar"
```

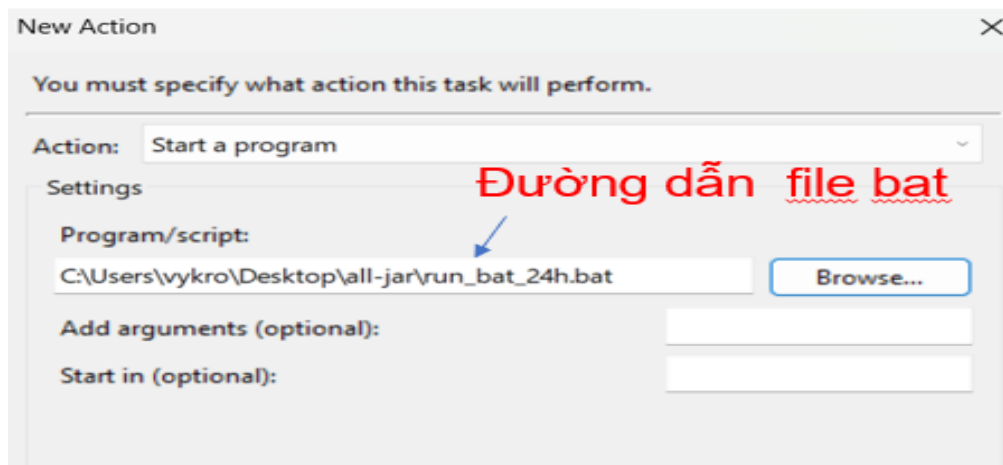
```
run_bat_24h.bat - Notepad
File Edit View
C:\Users\vykro\Desktop\all-jar\bin\run_cmd_24h.cmd
```

Bước 3: cài đặt Task Scheduler

Chọn mục Creat Task để thiết lập tự động hóa

The 'Create Task' dialog box is shown with the 'General' tab selected. It contains fields for 'Name:', 'Location:', 'Author:' (set to 'LAPTOPOFVYX\vykro'), and 'Description:'. A blue arrow points to the 'Name:' field with the red text 'Điền tên task' (Fill in the task name). At the bottom, there is a checkbox for 'Security options'.

The 'New Trigger' dialog box is shown with 'Begin the task:' set to 'On a schedule'. Under 'Settings', the 'Daily' radio button is selected. The 'Start:' date is '2/28/2023' and the time is '10:18:54 AM'. A blue arrow points to the time field with the red text 'Chọn thời gian chạy hàng ngày' (Choose daily running time). Below this, 'Recur every:' is set to '1' days. At the bottom, under 'Advanced settings', the 'Enabled' checkbox is checked. The 'OK' and 'Cancel' buttons are at the bottom right.



Sau khi hoàn thành các bước trên, máy tính theo đúng khung giờ đã cài đặt sẽ tự động thu thập dữ liệu và ghi vào file json với định dạng tên: 24h_<ngày thu thập>.json và express_<ngày thu thập>.json

V. Kết quả đạt được

- Chúng tôi đã xây dựng thành công một ứng dụng trích xuất thông tin từ các trang web tin tức 24h và vnexpress.
- Ứng dụng có thể tự động hóa việc thực hiện nhiệm vụ hàng ngày bằng ứng dụng Task Scheduler.
- Các thông tin bài báo được lưu trữ trong các file với định dạng .json để dễ dàng sử dụng và phân tích dữ liệu.

VI. Hướng phát triển tiếp theo:

- Mở rộng ứng dụng để trích xuất thông tin từ các trang web tin tức khác.
- Thêm tính năng lưu trữ các thông tin bài báo vào cơ sở dữ liệu để dễ dàng quản lý và sử dụng.
- Cải tiến giao diện người dùng để tăng tính thẩm mỹ
- Sử dụng server để chạy chương trình