**BÁO CÁO CÔNG VIỆC**

**Nhóm 2**

**1. Thành viên:**

* Bùi Nguyễn Nhật Tân
* Hồ Thị Kiều Linh
* Thái Thị Thanh Ngân
* Hạ Minh Chương
* Nguyễn Đình Mạnh Tiến

**1. Chủ đề:** Xây dựng website tổng hợp tin tức

**2.** **Ngôn ngữ sử dụng**: Java ( sử dụng thư viện hỗ trợ Jsoup)

**3. Phân công công việc thành viên nhóm**

|  |  |
| --- | --- |
| **Tên thành viên** | **Công việc** |
| Nguyễn Đình Mạnh Tiến | * Tìm hiểu các tool hỗ trợ cho việc crawl data |
| Thái Thị Thanh Ngân | * Tìm hiểu về khái niệm Crawl * Quy trình crawl dữ liệu |
| Bùi Nguyễn Nhật Tân | * Ưu – nhược điểm của Crawl dữ liệu từ web |
| Hồ Thị Kiều Linh | * Tìm hiểu về cách crawl dữ liệu từ trang web sử dụng tool hỗ trợ ngôn ngữ java |
| Hạ Minh Chương | * Lợi ích của Crawl, Mục đích của thu thập thông tin tự động crawl |

**I. Lý thuyết:**

**1.Khái niệm :**

- Web Crawling là một quá trình tự động duyệt qua các trang web và thu thập dữ liệu từ chúng. Dữ liệu này sau đó có thể được sử dụng cho nhiều mục đích khác nhau, chẳng hạn như tạo công cụ tìm kiếm, trang web so sánh giá và trang tổng hợp tin tức.

- Web Crawling Data có thể được trích xuất từ các loại trang web khác nhau, chẳng hạn như HTML, XML và JSON. Web Crawler, còn được gọi bot hay nhện, là một chương trình tự động điều hướng qua các trang web và trích xuất thông tin từ chúng.

- Crawling dữ liệu là quá trình tải về dữ liệu từ một số nguồn cụ thể trên internet, bao gồm các trang web, các trang mạng xã hội, các cổng dữ liệu, vv. Các thông tin được tải về có thể bao gồm các chi tiết như tên, địa chỉ, số điện thoại, email, vv.

**2. Quy trình crawl dữ liệu thường bao gồm các bước sau:**

   -Xác định mục tiêu: Xác định rõ mục tiêu và nguồn dữ liệu mà bạn muốn crawl.

   -Tìm kiếm và chọn nguồn dữ liệu: Tìm kiếm và chọn trang web, API hoặc cơ sở dữ liệu mà bạn muốn crawl dữ liệu từ.

   -Xác định cấu trúc dữ liệu: Sử dụng các công cụ phân tích HTML hoặc XML để xác định cấu trúc dữ liệu trên trang web mà bạn muốn crawl.

   - Viết mã để crawl dữ liệu: Sử dụng một ngôn ngữ lập trình, chẳng hạn như java hoặc Python, để viết mã để crawl dữ liệu từ trang web mục tiêu.

   - Chạy mã để crawl dữ liệu: Chạy mã để crawl dữ liệu và lưu trữ dữ liệu trong một tập tin hoặc cơ sở dữ liệu.

  -  Xử lý dữ liệu: Sử dụng các công cụ xử lý dữ liệu để xử lý và chuẩn hóa dữ liệu đã crawl.

   - Lưu trữ dữ liệu: Lưu trữ dữ liệu đã crawl trong một tập tin hoặc cơ sở dữ liệu và sẵn sàng để sử dụng

**3. Tìm hiểu các tool hỗ trợ cho việc crawl data**

Crawling dữ liệu có thể được sử dụng để xây dựng các cơ sở dữ liệu, phân tích dữ liệu cho mục đích kinh doanh, hoặc cho các mục đích nghiên cứu.

tool hỗ trợ cho việc thực hiện crawal data

Có rất nhiều công cụ hỗ trợ crawling dữ liệu từ internet. Một số trong số chúng ta có thể kể đến:

  +  Scrapy: là một framework Python mã nguồn mở dành cho việc crawling dữ liệu và lấy dữ liệu từ các trang web.

  +  BeautifulSoup: là một thư viện Python dành cho việc phân tích HTML và XML, cung cấp các công cụ cho việc lấy dữ liệu từ các trang web.

   + Octoparse: là một công cụ trực tuyến dành cho việc crawling dữ liệu, cho phép bạn tải về dữ liệu từ các trang web mà không cần biết lập trình.

   + Parsehub: là một công cụ trực tuyến cho việc crawling dữ liệu, cho phép bạn tải về dữ liệu từ các trang web với các tùy chọn tùy biến

   + Google Sheets: là một công cụ trực tuyến cho việc lưu trữ và quản lý dữ liệu, cung cấp các công cụ cho việc crawling dữ liệu từ internet với sử dụng các công thức.

**4. Ưu – nhược điểm**

Mặc dù Web Crawling có nhiều ứng dụng nhưng nó cũng có những bất lợi sau:

- Thu thập toàn bộ trang web có thể là một quá trình tốn nhiều thời gian và tài nguyên. Trình thu thập dữ liệu web phải được thiết kế để xử lý khối lượng dữ liệu lớn và có thể tăng hoặc giảm quy mô khi cần.

- Việc thu thập thông tin trên web có khả năng vi phạm quyền sở hữu trí tuệ hoặc vi phạm quyền riêng tư của người dùng.

- Web Crawling Data có thể chứa lỗi, trùng lặp hoặc thông tin không liên quan. Do đó, nó phải được thiết kế để lọc dữ liệu xấu và đảm bảo chất lượng dữ liệu.

**5. Cách crawl dữ liệu từ trang web sử dụng ngôn ngữ java**

Jsoup là Java HTML Parser. Nói cách khac Jsoup là một thư viện Java 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. Nó sử dụng các phương thức giống với DOM, CSS , JQuery để lấy dữ liệu và thao tác với dữ liệu.

Cụ thể, bạn có thể dùng Jsoup để lấy dữ liệu từ một trang web dạng code HTML, sau đó, bạn có thể dùng những hàm mà Jsoup cung cấp để lấy dữ liệu trong các class, các tag của file HTML đó. Dữ liệu bạn muốn lấy có thể là link một bức ảnh trong thẻ <img>, link một liên kết trong thẻ <a>, hay một đoạn text trong thẻ <p>,....

**Các hàm phổ biến của Jsoup**

Jsoup có ba class chính đó là :

* org.jsoup.Jsoup
* org.jsoup.nodes.Document
* org.jsoup.nodes.Element

1. **Class Jsoup:**

     Class Jsoup cung cấp những hàm để lấy Document từ một file HTML hoặc một URL, ví dụ hai hàm dưới đây :

     static Connection connect(String url).get(): Lấy Document của một trang web bằng URL

     static Document parse(String html) : Phân tích mã HTML và trả về đối tượng Document

**- Cài đặt Jsoup:** Bạn có thể tải Jsoup từ trang chủ của Jsoup hoặc thêm Jsoup vào Maven.

**- Phân tích HTML trang cần lấy:** Phân tích trang html tìm ra các đoạn code html đổ dữ liệu có thể lấy đưa vào database

- Sử dụng các hàm hỗ trợ của Jsoup tiến hành cào dữ liệu từ trang HTML

- Downlad source đã cào được hoặc đưa dữ liệu vào Database

**6.** **Lợi ích của Crawl, Mục đích của thu thập thông tin tự động crawl**

**Lợi ích của Crawl:**

**- Ít tốn thời gian và công sức trong quá trình lấy thông tin và dữ liệu:** Khi các bạn crawl dữ liệu, các bạn có một khối lượng thông tin cực kỳ lớn mà không phải tốn công nhập liệu.

**- Thao tác đơn giản dễ sử dụng:**Các bạn có thể lấy thông tin của nhiều website khác nhau để so sánh, đối chiếu và có thể làm thuật toán phân tích tiềm năng trên thị trường

- Đối với những người chơi affiliate thì thu thập thông tin tự động của các website bán hàng gần như là bắt buộc. Khi đó các bạn có thể liên hệ với công ty chúng tôi để có được sản phẩm ưng ý nhất được thiết kế trên nền web

**Mục đích của thu thập thông tin tự động crawl:**

**- Tăng lượt view cho các trang web để mọi người xung quanh biết đến:** Các bạn có thê thu thập thông tin tự động để phát triển website làm phong phú và đa dạng cho website của mình, thu hút người đọc một cách hiệu quả hơn và tăng lượt view cho mỗi trang web, giúp phát triển website một cách tốt nhất.

**- Đem lại hiệu quả cao trong công việc**: Các bạn muốn lấy thông tin, tin tức hoặc chính sạch mới nào đó của chính phủ hoặc luật pháp, thay vì phải vào website của họ copy nội dung đăng lên web của mình , các bạn hãy nghĩ ngay đến việc crawl dữ liệu, đó là phuong pháp vừ nhanh, không mất thời gian, hoàn toàn tự động và chính xác nhất.