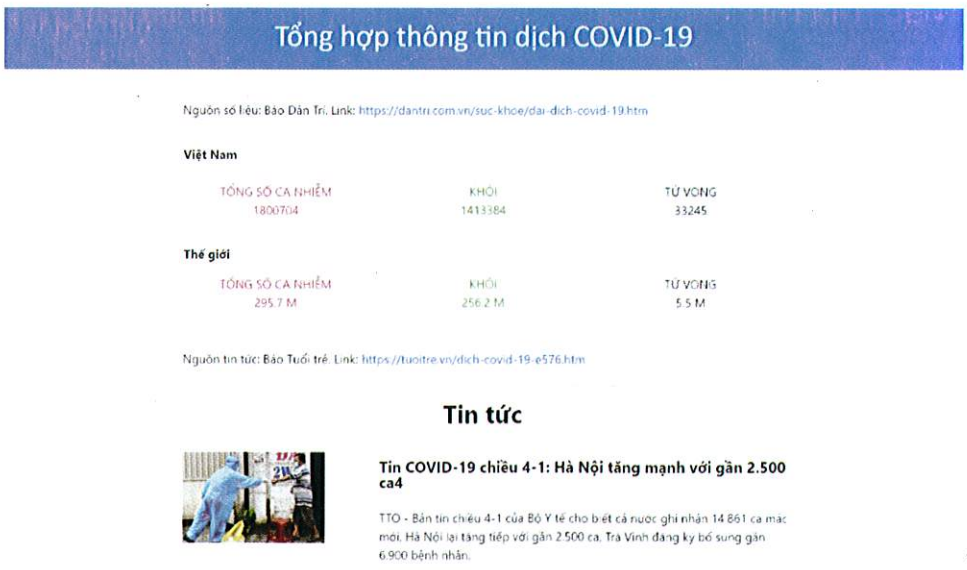
CHƯƠNG IV

Lập trình web

Trong chương này, chúng ta sẽ tạo một hệ thống website giúp thu thập tin tức về dịch Covid-19 từ các trang báo điện tử. Các bạn sẽ được tìm hiểu cách hoạt động của một hệ thống website, làm thế nào mà các trình duyệt như Google Chrome, Microsoft Edge, Cốc cốc... có thể hiển thị nội dung website hay cách xây dựng một website bằng Python. Hành động thu thập thông tin này có từ khóa là Web Scraping.

Chúng ta sẽ trải qua các bước:

- Tìm hiểu về quá trình truy cập một website.
- Tìm hiểu cấu trúc các tập tin thành phần của website và cách xây dựng một website sử dụng Django framework.
- Tìm hiểu cách trình duyệt hiển thị website bằng ngôn ngữ đánh dấu siêu văn bản HTML.
- Lập trình thu thập thông tin và tin tức về dịch Covid-19 trên các trang báo điện tử có định dạng mã HTML.
- Tách các dữ liệu từ mã HTML và hiển thị lên giao diện website.



BÀI 1 CẤU TRÚC WEBSITE

Đã bao giờ các bạn tự mình đặt câu hỏi: Một website hoạt động như thế nào? Khi truy cập vào một website bất kỳ như https://www.youtube.com/, dữ liệu được lấy từ đâu? Trình duyệt web hiển thị các dữ liệu bằng cách nào? Bài học này sẽ cung cấp những khái niệm và kiến thức nền tảng để các bạn nắm được cách một website bất kỳ hoạt động.

1.1. Một số khái niệm cơ bản

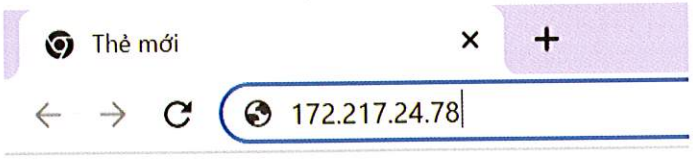
Để hiểu được cách một website hoạt động, chúng ta cần hiểu được đúng các khái niệm sau:

Siêu văn bản: Dạng văn bản tích hợp nhiều loại dữ liệu khác nhau như văn bản, hình ảnh, âm thanh, video… và các liên kết đến các siêu văn bản khác. Thông tin trên internet thường được tổ chức dưới dạng các siêu văn bản. Các siêu văn bản thường được hiển thị bằng ngôn ngữ đánh dấu siêu văn bản HTML HyperText Markup Language (được giới thiệu chi tiết trong Bài 5) và là thành phần chính của các trang web.

Website: Hệ thống một hoặc nhiều trang web có liên quan và liên kết đến nhau được tổ chức thành website. Mỗi website có một địa chỉ truy cập nhất định và được lưu trữ trên các máy chủ. Khi ta truy cập một website, trang web được mở ra đầu tiên được gọi là trang chủ (Homepage).

Trình duyệt web: Chương trình được sử dụng để truy cập các website và hiển thị nội dung trang web. Một số trình duyệt web phổ biến hiện nay là Cốc cốc, Google Chrome, Microsoft Edge, Safari, Firefox…

Địa chỉ IP: Mã gồm các chữ số được dùng để định danh cho một website với giao thức internet (IP – Internet Protocol), với IP phiên bản 4 (IPv4) địa chỉ là một số 32-bit, ví dụ 157.240.211.35 là địa chỉ của Facebook, một số website có thể có nhiều địa chỉ IP khác nhau với mục đích cân bằng tải. Chúng ta có thể sử dụng trình duyệt truy cập trực tiếp đến địa chỉ IP này.



Hãy thử xem địa chỉ 172.217.24.78 là của website nào?

Tên miền (domain): Để truy cập vào các website, trình duyệt cần địa chỉ IP là dãy các con số, tuy nhiên đối với con người địa chỉ IP rất khó nhớ và không thuận tiện sử dụng, nên ta "đặt tên" cho các địa chỉ IP này để dễ nhớ và dễ sử dụng hơn, đó là tên

miền. Tên miền là địa chỉ của website nhưng thể hiện dưới dạng chuỗi ký tự văn bản, ví dụ tên miền của Google là "google.com".

DNS (Domain Name System): Hệ thống phân giải tên miền, giúp chuyển đổi tên miền của website thành địa chỉ IP tương ứng.

Máy chủ web (web server): Máy tính vật lý chứa dữ liệu của website các như tệp tin, hình ảnh, âm thanh, thông tin người dùng…

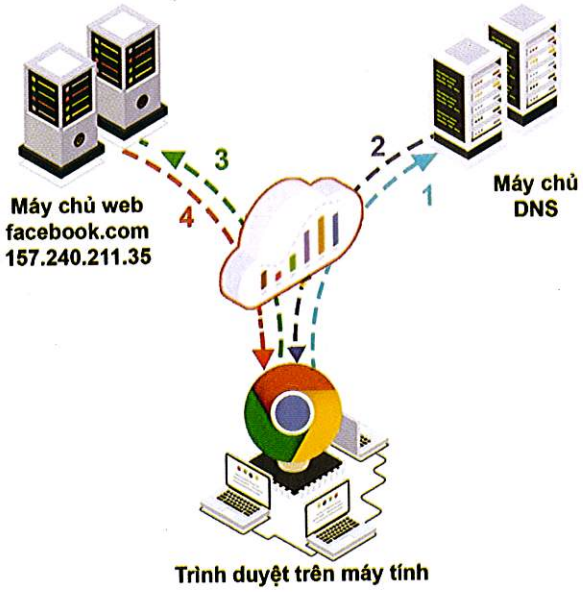
Hosting: Không gian trên máy chủ có cài dịch vụ Internet, nơi lưu trữ nội dung hay dữ liệu của website trên máy chủ.

HTML: Ngôn ngữ đánh dấu văn bản, là ngôn ngữ dùng để hiển thị các siêu văn bản trên trình duyệt

Cơ sở dữ liệu: Hệ thống các thông tin, dữ liệu có tổ chức.

1.2. Sơ đồ hoạt động website

Khi chúng ta sử dụng trình duyệt để truy cập một website thông qua tên miền (ví dụ ta gõ "facebook.com" và nhấn nút Enter), các công việc được thực hiện tiếp sau đó được tóm tắt và đơn giản hóa theo sơ đồ sau:



1. Trình duyệt gửi yêu cầu phân giải địa chỉ đến máy chủ DNS để xem tên miền "facebook.com" có địa chỉ IP là gì.
2. Máy chủ DNS gửi phản hồi địa chỉ IP của tên miền "facebook.com", ví dụ 157.240.211.35
3. Trình duyệt dựa trên địa chỉ IP nhận được để gửi yêu cầu HTTP (HTTP Request) đến máy chủ website với yêu cầu truy xuất nội dung hiển thị trang web facebook.com

4. Máy chủ của Facebook sẽ xử lý và gửi trả lời yêu cầu này, trong đó có những tệp tin HTML giúp trình duyệt hiển thị giao diện của trang web facebook.com lên máy tính. Máy chủ của Facebook sẽ liên tục lắng nghe xem có yêu cầu nào được gửi đến không và trả lời các yêu cầu này.

Trong Chương 4 này, chúng ta sẽ làm dự án là tạo ra một trang web có chức năng tổng hợp các thông tin về dịch Covid-19, đồng thời chúng ta cũng xây dựng một hệ thống với vai trò như một máy chủ web để vận hành.

Tóm tắt lý thuyết và bài tập thực hành

Trong bài học này, chúng ta đã được tìm hiểu về một số khái niệm cơ bản về website và sơ đồ hoạt động của một website bất kỳ. Các bạn có thể tự tìm hiểu thêm các thông tin dựa trên những từ khóa trong cuốn sách cung cấp.

Bài tập 1. Em hãy tự mình vẽ lại sơ đồ hoạt động và luồng xử lý khi một máy tính bất kỳ truy cập vào trang chủ https://vietstem.com/.

Bài tập 2. Máy chủ có chức năng gì?

Bài tập 3. Website và trang web khác nhau như thế nào?