**HỌC CÁCH CRAWL SỐ LIỆU VỀ DỊCH BỆNH COVIDE-19**

**TỪ WEBSITE WORDOMETER BẰNG PHP**

**1. Giới thiệu về website Worldometer, và trang thông tin về dịch bệnh COVID-19.**

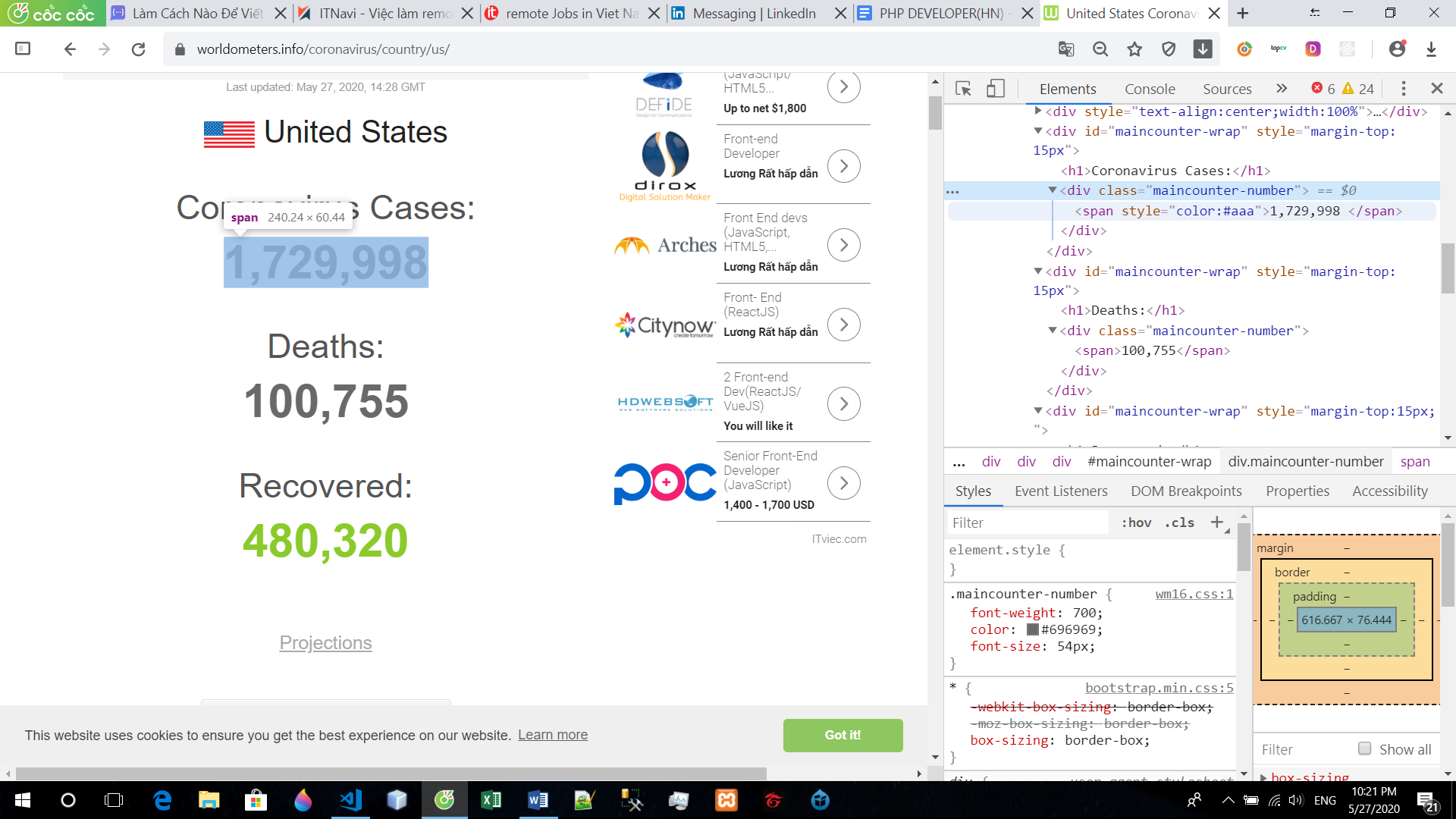
**Worldometer** [[2]](https://vi.wikipedia.org/wiki/Worldometer#cite_note-NameFAQ-2) là một [trang web](https://vi.wikipedia.org/wiki/Website) tham khảo cung cấp số liệu [thống kê](https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA) thời gian thực cho các chủ đề đa dạng. Website được sáng lập vào năm 2007 và điều hành bởi một công ty dữ liệu của Mỹ, Dadax, được [Hiệp hội Thư viện Hoa Kỳ](https://vi.wikipedia.org/wiki/Hi%E1%BB%87p_h%E1%BB%99i_Th%C6%B0_vi%E1%BB%87n_Hoa_K%E1%BB%B3) bầu chọn là một trong những trang web tham khảo miễn phí tốt nhất (năm 2011)

Năm 2020, đại dịch Covid-19 bùng nổ, Worldometer đã cho mở chuyên trang về cập nhập tình hình diễn biến của dịch bệnh trên toàn cầu, và theo từng quốc gia. Và chuyên trang này đã giúp worldometer trở nên nổi tiếng và được sử dụng rộng rãi trên toàn thế giới.

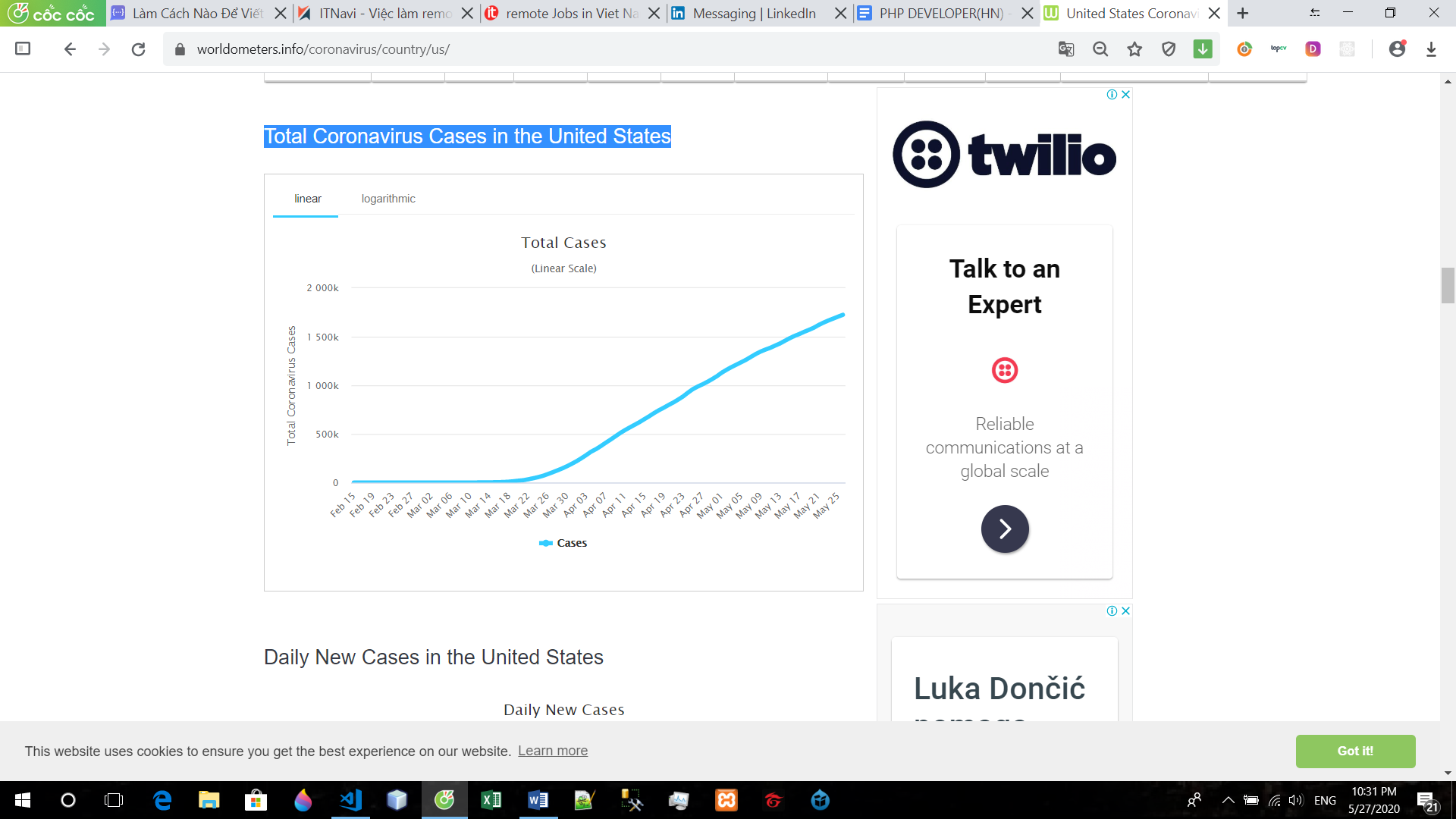
**2. Phân tích đặc điểm bố cục trang thông tin dịch bệnh COVID-19 của Worldometer.**

Để crawl được dữ liệu từ một trang web, ta cần phân tích bố cục, cách trình bày dữ liệu của trang web đó. Căn cứ vào đó để có phương pháp crawl phù hợp nhất.

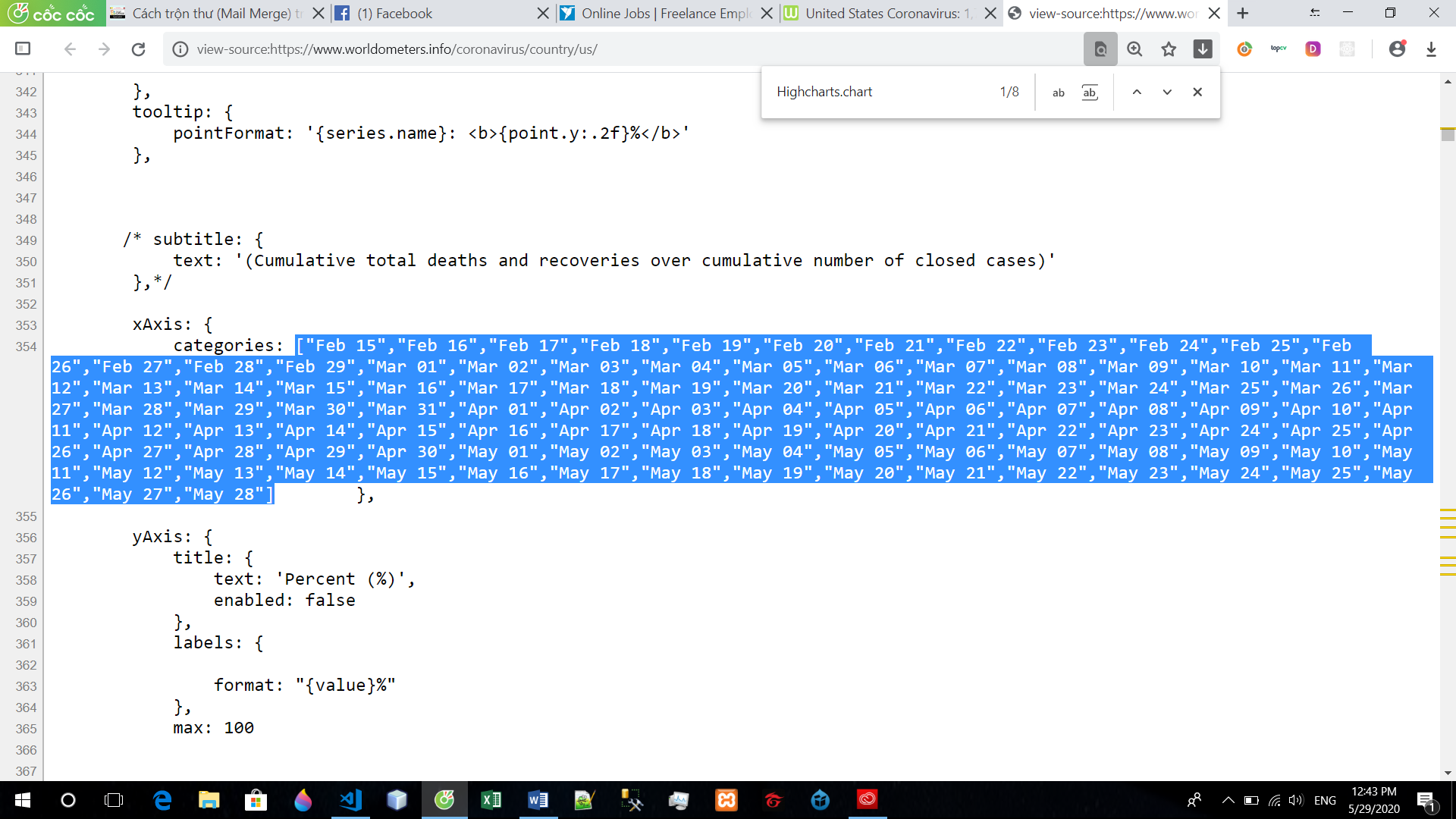
Hãy cùng phân tích trang web dưới đây để biết Worldometer cách trình bày số liệu về dịch bệnh Covid-19 tại một quốc gia cụ thể: <https://www.worldometers.info/coronavirus/country/us/> .  
 Đường link trên dẫn tới một trang web trình bày số liệu thống kê dịch bệnh Covid-19 tại Mỹ:

Phân tích cấu trúc của trang web trên ta thấy, các số liệu về số ca nhiễm, số ca tử vong, số ca được chữa khỏi đều được hiển thị trong thẻ *<span>* là child của thẻ *<div class=”maincounter-number”>*. Điều đó cho phep ta lấy các số liệu đó thông qua *dom selector “.maincounter-number span”*  


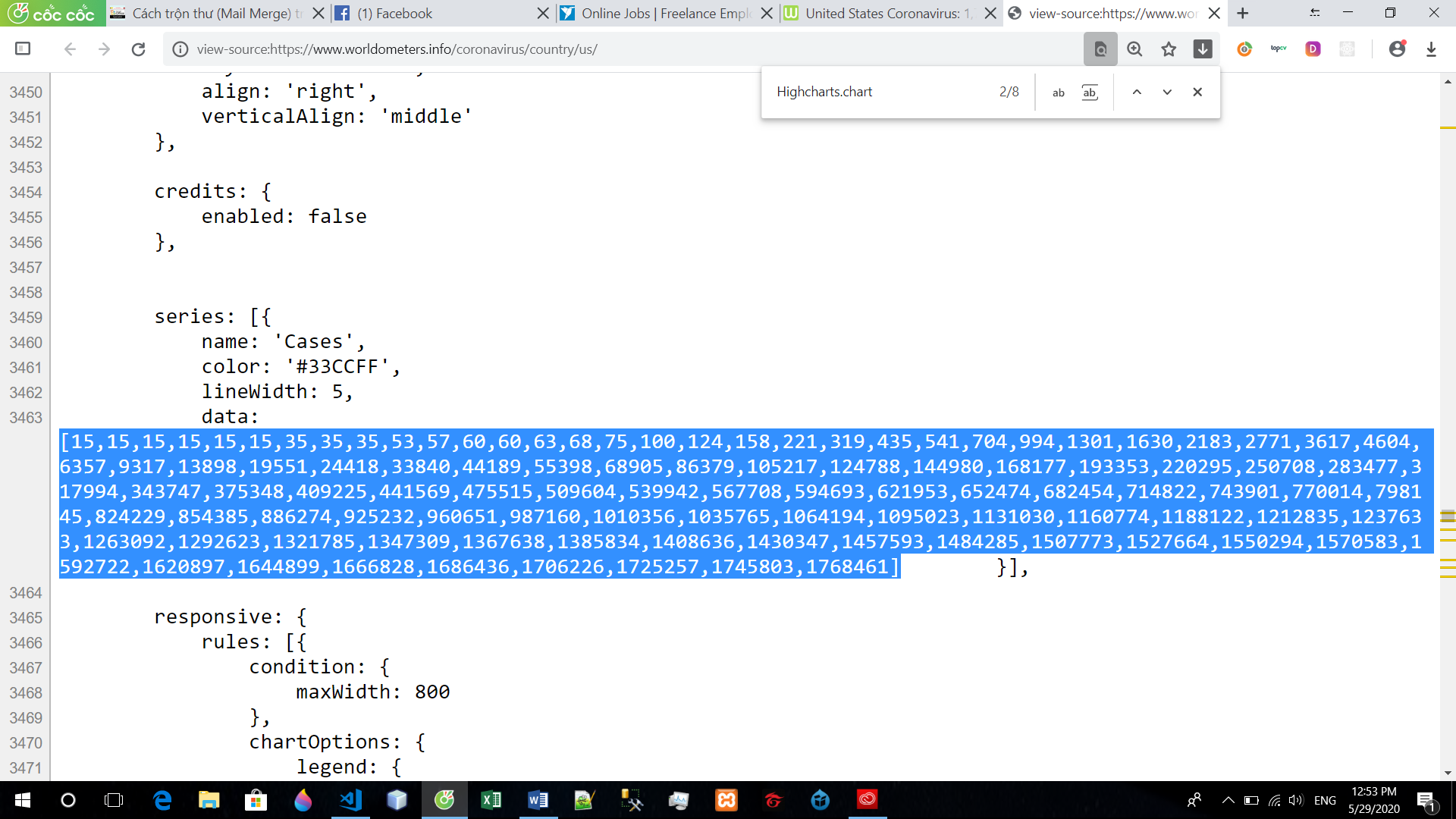
Về tập số liệu miêu tả diễn biến số ca nhiễm Covid-19 qua từng ngày, ta có thể thấy Worldometer biểu diễn nó dưới dạng biểu đồ.



Tra cứu từ khóa “*coronavirus-cases-linear*” trong mã nguồn của Worldometer, ta có thể thấy Worldometer đã sử dụng framework Highchart.js để xây dựng biểu đồ trên. Và Worldometer đã cung cấp cho framework 02 mảng dữ liệu để hiển thị ngoài output. Một mảng là *xAxis* thể hiện các mốc thời gian trên trục hoành (chính là các mốc thời gian từ “Feb 15” đến nay), một mảng khác là *data*, thể hiện danh sách số người mắc Covid-19 qua các ngày từ “Feb 15” đến nay.



*Mảng xAxis thể hiện các mốc thời gian*



*Mảng data danh sách số người mắc (Case) Covid-19 qua các mốc thời gian*

Tương tự, ta tiếp tục tra cứu với các từ khóa *“graph-cases-daily”, “graph-active-cases-total”, “coronavirus-deaths-linear”, “graph-deaths-daily”,* ta sẽ có được cách dữ liệu tương tự về *số ca nhiễm mới, số ca nhiễm trên thực tế, số ca tử vong,* và *số ca tử vong mới*  theo các mốc thời gian (ngày) cụ thể*.*

**3. Viết một chương trình crawler.**

**a. Giới thiệu chương trình (nội dung, mục đích, công nghệ, cách hoạt động).**

**b. Giới thiệu thư viện Simple Html Dom của PHP.**

**c. Các bước thực hiện xây dựng chương trình Crawler theo mô hình MVC.**