Tóm tắt mô-đun 5: API và thu thập dữ liệu

Chúc mừng! Bạn đã hoàn thành mô-đun này. Tại thời điểm này, bạn biết rằng:

- API đơn giản trong Python là các giao diện lập trình ứng dụng cung cấp các phương thức đơn giản và dễ sử dụng để tương tác với các dịch vụ, thư viện hoặc dữ liệu, thường với cấu hình hoặc độ phức tạp tối thiểu.
    - API cho phép hai phần mềm nói chuyện với nhau.
    - Sử dụng thư viện API trong Python đòi hỏi phải nhập thư viện, gọi các hàm hoặc phương thức của nó để thực hiện các yêu cầu HTTP và phân tích cú pháp các phản hồi để truy cập dữ liệu hoặc dịch vụ do API cung cấp.
    - Pandas API xử lý dữ liệu bằng cách giao tiếp với các thành phần phần mềm khác.
    - Phiên bản hình thành khi bạn tạo từ điển và sau đó sử dụng hàm tạo DataFrames để tạo đối tượng Pandas.
    - Phương thức "head()" sẽ hiển thị số hàng được đề cập từ trên cùng (mặc định 5) của DataFrames, trong khi phương thức "mean()" sẽ tính giá trị trung bình và trả về các giá trị
- Các API còn lại cho phép bạn giao tiếp thông qua internet, tận dụng các tài nguyên như lưu trữ, truy cập nhiều dữ liệu hơn, thuật toán AI, v.v.
    - Các phương thức HTTP truyền dữ liệu qua internet.
    - Thông báo HTTP thường bao gồm tệp JSON với hướng dẫn hoạt động.
    - Thông điệp HTTP chứa tệp JSON được trả về máy khách dưới dạng phản hồi từ các dịch vụ web.
    - Xử lý dữ liệu chuỗi thời gian liên quan đến việc sử dụng hàm chuỗi thời gian Pandas.
    - Bạn có thể lấy dữ liệu cho nến hàng ngày và vẽ biểu đồ bằng Plotly với biểu đồ nến.
- HTTP (Giao thức truyền siêu văn bản) truyền dữ liệu, bao gồm các trang web và tài nguyên, giữa máy khách (trình duyệt web) và máy chủ trên World Wide Web.
    - Giao thức HTTP thường được sử dụng để triển khai các loại API REST khác nhau.
    - Phản hồi HTTP bao gồm thông tin như loại tài nguyên, độ dài của tài nguyên, v.v
    - Bộ định vị tài nguyên thống nhất (URL) là cách phổ biến nhất để tìm tài nguyên trên web.
    - URL được chia thành ba phần: lược đồ, địa chỉ internet hoặc URL cơ sở và tuyến đường
    - Phương thức GET là một trong những phương pháp yêu cầu thông tin phổ biến. Một số phương pháp khác cũng có thể bao gồm cơ thể.
    - Phương thức phản hồi chứa phiên bản và nội dung của phản hồi.
    - POST gửi dữ liệu lên máy chủ, PUT cập nhật dữ liệu đã có trên máy chủ, DELETE xóa dữ liệu khỏi máy chủ
- Yêu cầu là một thư viện Python cho phép bạn gửi các yêu cầu HTTP / 1.1 một cách dễ dàng
    - Bạn có thể sửa đổi kết quả truy vấn của mình bằng phương thức GET.
    - Bạn có thể nhận nhiều yêu cầu từ một URL như tên, ID, v.v. bằng chuỗi Truy vấn.
- Web scraping trong Python liên quan đến việc trích xuất và phân tích dữ liệu từ các trang web để thu thập thông tin cho các ứng dụng khác nhau, sử dụng các thư viện như Beautiful Soup và các yêu cầu.
    - HTML bao gồm văn bản được bao quanh bởi các phần tử văn bản màu xanh lam được đặt trong dấu ngoặc góc được gọi là thẻ.
    - Bạn có thể chọn một phần tử HTML trên một trang web để kiểm tra trang web.
    - Các trang web cũng có thể chứa CSS và JavaScript cùng với các phần tử HTML.
    - Mỗi tài liệu HTML giống như một cây HTML, có thể chứa chuỗi và các thẻ khác.
    - Mỗi bảng HTML bao gồm các thẻ bảng và được cấu trúc với các phần tử như hàng, tiêu đề, nội dung, v.v.
- Dữ liệu dạng bảng cũng có thể được trích xuất từ các trang web bằng phương pháp 'read_html' trong Pandas.
- Beautiful Soup trong Python là một thư viện để phân tích cú pháp và điều hướng các tài liệu HTML và XML, giúp trích xuất và thao tác dữ liệu từ các trang web dễ tiếp cận hơn.
- Để phân tích cú pháp một tài liệu, hãy truyền nó qua hàm tạo Beautiful Soup để có được một đối tượng súp đẹp đại diện cho tài liệu dưới dạng cấu trúc dữ liệu lồng nhau.
- Súp đẹp đại diện cho HTML như một tập hợp các đối tượng giống như cây với các phương thức để phân tích cú pháp HTML.
- Chuỗi có thể điều hướng giống như một chuỗi Python hỗ trợ chức năng súp đẹp mắt.
- find_all là một phương pháp được sử dụng để trích xuất nội dung dựa trên tên của thẻ, các thuộc tính của nó, văn bản của một chuỗi hoặc một số kết hợp của chúng.
- Phương thức find_all xem xét các hậu duệ của thẻ và truy xuất tất cả các hậu duệ phù hợp với bộ lọc của bạn.
- Kết quả là một Python có thể lặp lại như một danh sách.
- Định dạng tệp đề cập đến cấu trúc và quy tắc mã hóa cụ thể được sử dụng để lưu trữ và biểu diễn dữ liệu trong tệp, chẳng hạn như .txt cho văn bản thuần túy hoặc .csv cho các giá trị được phân tách bằng dấu phẩy.
- Python hoạt động với các định dạng tệp khác nhau như CSV, XML, JSON, xlsx, v.v
- Phần mở rộng của tên tệp sẽ cho bạn biết loại tệp đó là gì và nó cần mở bằng gì.
- Để truy cập dữ liệu từ các tệp CSV, chúng ta có thể sử dụng các thư viện Python như Pandas.
- Tương tự, các phương pháp khác nhau giúp phân tích cú pháp JSON, XML và các tệp khác.

Chuyển đến mục tiếp theo ✓ **Hoàn thành**