Machine Translated by Google

Rút trích nội dung trang web

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Quét web là một thuật ngữ chung cho kỹ thuật liên quan đến việc tự động hóa việc thu thập dữ liệu từ một trang web.
- Trong phần này, chúng ta sẽ tìm hiểu cách sử dụng Python để thực hiện các tác vụ quét web, chẳng hạn như tải xuống hình ảnh hoặc thông tin từ một trang web.

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Để quét web bằng Python, chúng tôi cần hiểu các khái niệm cơ bản về cách hoạt động của một trang web.
- Khi trình duyệt tải một trang web, người dùng sẽ thấy phần được gọi là “giao diện người dùng” của trang web.

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

Machine Translated by Google

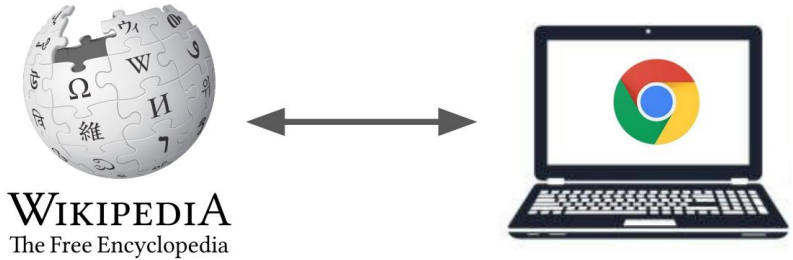
Hoàn thành chương trình đào tạo Python

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

Machine Translated by Google

Hoàn thành chương trình đào tạo Python



Machine Translated by Google

Hoàn thành chương trình đào tạo Python

PIERIAN DATA

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

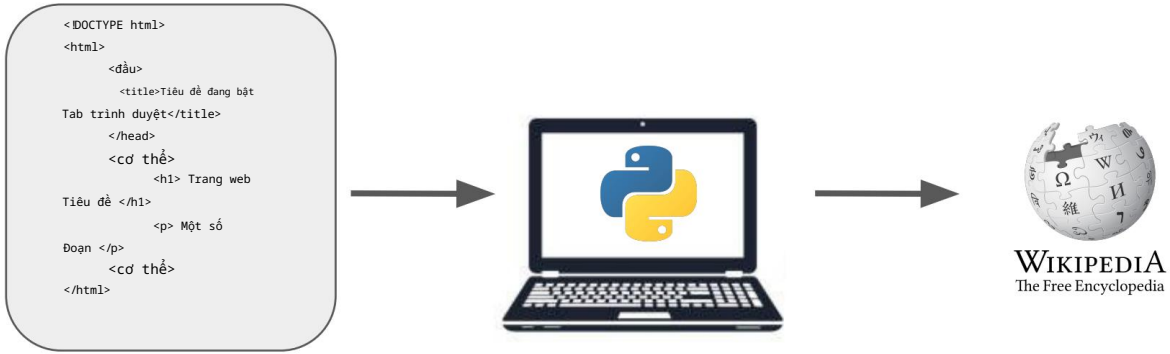
```
<!DOCTYPE html> <html>

      <đầu>
        <title>Tiêu đề đang bật
Tab trình duyệt</title>
      </head>
      <cơ thể>
            <h1> Trang web
Tiêu đề </h1>
            <p> Một số
Đoạn </p>
      <cơ thể>
</html>
```

PIERIAN DATA

Machine Translated by Google

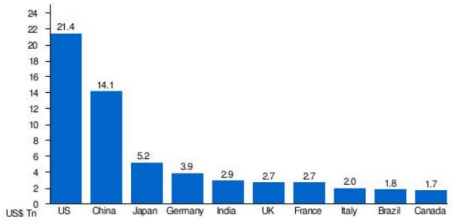
Hoàn thành chương trình đào tạo Python



Machine Translated by Google

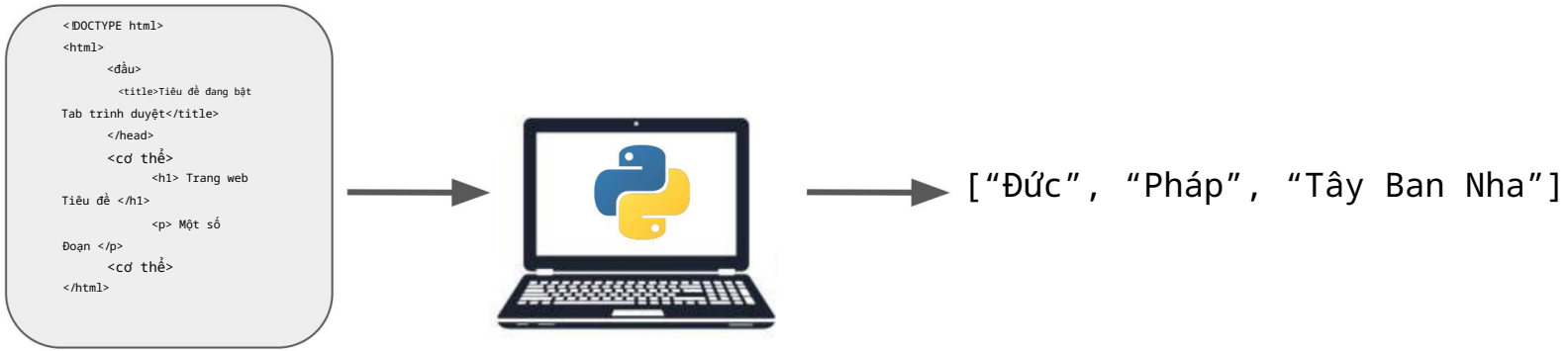
Hoàn thành chương trình đào tạo Python

```
<!DOCTYPE html>
<html>
      <đầu>
        <title>Tiêu đề đang bật
Tab trình duyệt</title>
      </head>
      <cơ thể>
            <h1> Trang web
Tiêu đề </h1>
            <p> Một số
Đoạn </p>
      <cơ thể>
</html>
```



Machine Translated by Google

Hoàn thành chương trình đào tạo Python



Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Những điều chính chúng ta cần hiểu • Quy tắc quét web • Hạn chế của quét web • HTML và CSS cơ bản

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Quy tắc quét web
  - Luôn cố gắng xin phép trước khi cạo!
  - Nếu bạn cạo quá nhiều các lần thử hoặc yêu cầu Địa chỉ IP của bạn có thể bị chặn!
  - Một số trang web tự động chặn phần mềm thu thập dữ liệu.

Machine Translated by Google

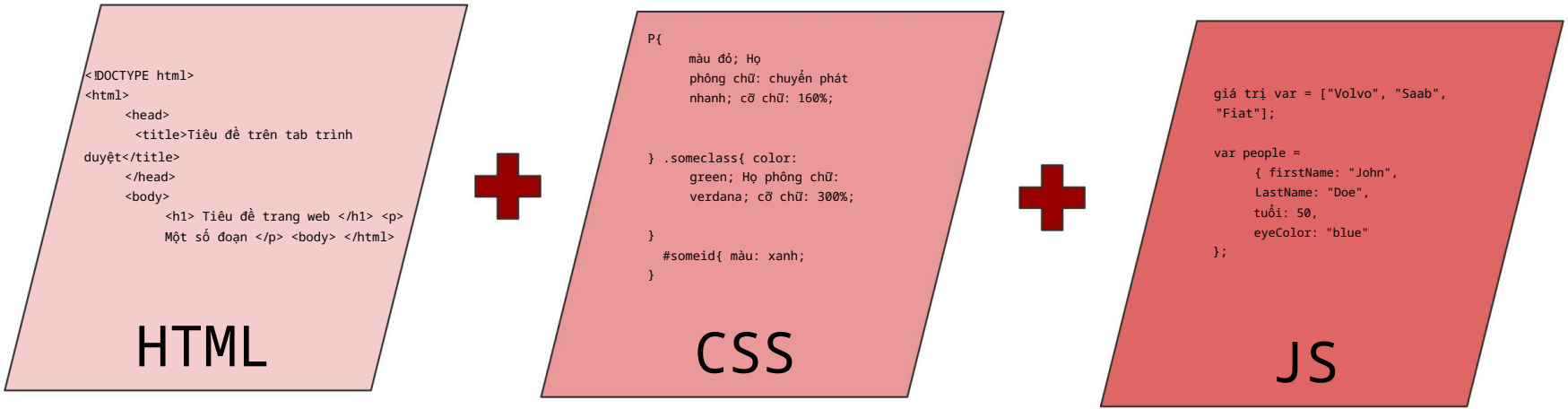
Hoàn thành chương trình đào tạo Python

- Hạn chế của việc quét web
  - Nói chung, mỗi trang web là duy nhất, có nghĩa là mỗi tập lệnh quét web là duy nhất.
  - Một thay đổi nhỏ hoặc cập nhật cho một trang web có thể phá vỡ hoàn toàn tập lệnh quét web của bạn.

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

Các thành phần giao diện người dùng chính của một trang web



PIERIAN DATA

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Khi xem một trang web, trình duyệt không hiển thị cho bạn tất cả mã nguồn đằng sau trang web, thay vào đó nó hiển thị cho bạn HTML và một số CSS và JS mà trang web gửi tới trình duyệt của bạn.

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- HTML được sử dụng để tạo cấu trúc và nội dung cơ bản của trang web
- CSS được sử dụng để thiết kế và tạo kiểu cho một trang web, nơi các phần tử được đặt và nó trông như thế nào
- JavaScript được sử dụng để xác định tính tương tác các thành phần của một trang web

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Để quét web cơ bản hiệu quả, chúng ta chỉ cần có hiểu biết cơ bản về HTML và CSS.

- Python có thể xem các phần tử HTML và CSS này theo chương trình, sau đó trích xuất thông tin từ trang web.
- Hãy khám phá HTML và CSS chi tiết hơn.

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- HTML là Ngôn ngữ đánh dấu siêu văn bản và có mặt trên mọi trang web trên internet.
- Bạn có thể nhấp chuột phải vào trang web và chọn “Xem nguồn trang” để xem ví dụ.
- Hãy xem một ví dụ nhỏ về mã HTML.

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

```
<!DOCTYPE html>
<html>
      <head>
        <title>Tiêu đề trên tab trình duyệt</title> </
      head>
      <body>
            <h1> Tiêu đề trang web </h1> <p>
            Một số đoạn </p> <body> </html>
```

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

```
<!DOCTYPE html>
<html>
      <head>
        <title>Tiêu đề trên tab trình duyệt</title> </
      head>
      <body>
           <h1> Tiêu đề trang web </h1> <p>
           Một số đoạn </p> <body> </html>
```

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

```
<!DOCTYPE html>
<html>
      <head>
       <title>Tiêu đề trên tab trình duyệt</title> </
      head>
      <body>
            <h1> Tiêu đề trang web </h1> <p>
            Một số đoạn </p> <body> </html>
```

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

```
<!DOCTYPE html>
<html>
      <head>
        <title>Tiêu đề trên tab trình duyệt</title> </
      head>
      <body>
           <h1> Tiêu đề trang web </h1> <p>
           Một số đoạn </p> <body> </html>
```

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- CSS là viết tắt của Cascading Style Sheets.
- CSS mang lại “phong cách” cho trang web, chẳng hạn như thay đổi màu sắc và phông chữ.
- CSS sử dụng thẻ để xác định phần tử html nào sẽ được tạo kiểu.

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

```
<!DOCTYPE html>
    <html>
        <head>
      <link rel="stylesheet" href="styles.css"> <title>Một
      số tiêu đề</title> </head> <body>
      <p

      id='para2'> Một số văn bản </p> <body> </
      html>
```

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

```
<!DOCTYPE html>
    <html>
        <head>
        <link rel="stylesheet" href="styles.css"> <title>Một
        số tiêu đề</title> </head> <body>
        <p

        id='para2'> Một số văn bản </p> <body> </
        html>
```

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

```
<!DOCTYPE html>
    <html>
        <head>
        <link rel="stylesheet" href="styles.css"> <title>Một
        số tiêu đề</title> </head> <body>
        <p

        id='para2'> Một số văn bản </p> <body> </
        html>
```

Machine Translated by Google

Ví dụ về tệp style.css:

```
#para2
      { màu: đỏ;
}
```

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

```
<!DOCTYPE html> <html>

  <đầu>

<link rel="stylesheet" href="styles.css"> <title>Một số
tiêu đề</title> </head> <body> <p

class='cool'> Một số văn bản </p> <body> </ html>
```

Machine Translated by Google

Ví dụ về tệp style.css:

```
.cool
    { màu: đỏ; Họ
    phông chữ: verdana;
}
```

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

```
P{
     màu đỏ; Họ
     phông chữ: chuyển phát
     nhanh; cỡ chữ: 160%;



} .someclass{ color:
     green; Họ phông chữ:
     verdana; cỡ chữ: 300%;

}
  #someid{ màu: xanh;
}
```

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Đừng lo lắng về việc ghi nhớ điều này! Chúng ta sẽ thấy rất nhiều ví dụ, những ý chính cần lưu ý:
  - HTML chứa thông tin
  - CSS chứa kiểu dáng
  - Chúng ta có thể sử dụng thẻ HTML và CSS để định vị thông tin cụ thể trên một trang

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Để thu thập dữ liệu web bằng Python, chúng ta có thể sử dụng thư viện BeautifulSoup và yêu cầu.
- Đây là các thư viện bên ngoài bên ngoài Python nên bạn cần cài đặt chúng bằng conda hoặc pip tại dòng lệnh của mình.

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Trực tiếp tại dòng lệnh của bạn, hãy sử dụng:
  - yêu cầu cài đặt pip
  - cài đặt pip lxml
  - cài đặt pip bs4
- Hoặc đối với bản phân phối Anaconda, hãy sử dụng cài đặt conda thay vì cài đặt pip.

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Hãy cùng xem qua một số ví dụ về web cào bằng Python!

Machine Translated by Google

Đang cài đặt
Để quét web

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Cài đặt các thư viện cần thiết
- Khám phá cách kiểm tra các thành phần và xem nguồn của trang web
- Lưu ý: Chúng tôi khuyên bạn nên sử dụng Chrome để bạn có thể làm theo chính xác như chúng tôi, nhưng những công cụ này có sẵn trong tất cả các trình duyệt chính.

Machine Translated by Google

Lấy tiêu đề trang

Machine Translated by Google

Nắm lấy tất cả
Các phần tử của một lớp

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Trước đây chúng tôi đã đề cập đến một phần quan trọng của việc tìm kiếm web bằng thư viện BeautifulSoup là tìm ra cú pháp chuỗi nào cần chuyển vào phương thức soup.select().
- Chúng ta hãy xem qua một bảng với một số các ví dụ phổ biến (những ví dụ này rất có ý nghĩa nếu bạn biết cú pháp CSS)

Machine Translated by Google

Cú pháp	Kết quả trận đấu
súp.select('div')	Tất cả các phần tử có thẻ 'div'
súp.select('#some_id')	Các phần tử chứa id='some_id'
súp.select('.some_class')	Các phần tử chứa class = 'some_class'
súp.select('div span')	Bất kỳ phần tử nào có tên span trong phần tử div.
súp.select('div > span')	Bất kỳ phần tử nào được đặt tên span trực tiếp trong phần tử div, không có phần tử nào ở giữa.

Machine Translated by Google

Lấy một hình ảnh

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Bây giờ chúng ta đã hiểu cách lấy thông tin văn bản dựa trên thẻ và tên thành phần, hãy khám phá cách lấy hình ảnh từ một trang web.

- Hình ảnh trên trang web thường có liên kết URL riêng (kết thúc bằng .jpg hoặc .png)

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Beautiful Soup có thể quét một trang, định vị các thẻ <img> và lấy các URL này.
- Sau đó, chúng tôi có thể tải các URL xuống dưới dạng hình ảnh và ghi chúng vào máy tính.
- Lưu ý: Bạn phải luôn kiểm tra bản quyền quyền trước khi tải xuống và sử dụng hình ảnh từ một trang web.

Machine Translated by Google

Làm việc với Nhiều trang và mục

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Chúng tôi đã biết cách lấy từng phần tử một nhưng trên thực tế, chúng tôi muốn có thể lấy nhiều phần tử, rất có thể là trên nhiều trang.

- Đây là nơi chúng ta có thể kết hợp các kiến thức về python với các thư viện quét web để tạo các tập lệnh mạnh mẽ!

Machine Translated by Google

Hoàn thành chương trình đào tạo Python

- Chúng tôi sẽ sử dụng một trang web được thiết kế đặc biệt để thực hành quét web: www.toscrape.com
- Chúng tôi sẽ thực hành lấy các phần tử trên nhiều trang.
- Hãy bắt đầu!

Machine Translated by Google

Làm việc với Nhiều trang

Machine Translated by Google

Bài tập quét web

Tổng quan

Machine Translated by Google

Rút trích nội dung trang web

Bài tập Giải pháp

Machine Translated by Google

Bài tập quét web

Giải pháp - Phần thứ hai