## 1.Giới thiệu bài toán

Bài toán cào dữ liệu ảnh (Image Scrapping) là quá trình tự động trích xuất thông tin từ các hình ảnh trên Internet. Đây là một công việc quan trọng trong việc thu thập dữ liệu từ web để phân tích, khai thác thông tin hoặc huấn luyện mô hình trí tuệ nhân tạo.

Việc cào dữ liệu ảnh được sử dụng rộng rãi trong nhiều lĩnh vực như bán lẻ, y tế, giáo dục, nghiên cứu khoa học, truyền thông và giải trí. Ví dụ, các công ty bán lẻ có thể sử dụng dữ liệu cào để tìm kiếm thông tin về sản phẩm và giá cả của đối thủ cạnh tranh, còn các nhà nghiên cứu y học có thể sử dụng dữ liệu cào để phân tích hình ảnh từ các bệnh nhân để tìm ra các biểu hiện của bệnh lý.

Chương trình sử dụng thư viện web scrapping rất phổ biến BeautifulSoup để lấy dữ liệu ảnh của Google Images:

1. Tải nội dung HTML của trang web chứa hình ảnh bằng cách sử dụng thư viện ‘request’ của Python.

2. Sử dụng BeautifulSoup để phân tích cú pháp HTML và tìm kiếm tất cả các phần tử ‘img’ trên Google Images.

3. Trích xuất các thông tin cần thiết từ các phần tử ‘img’, chẳng hạn như đường dẫn URL và tên tệp, sau đó lưu chúng vào một danh sách hoặc tệp dữ liệu.

## 2. Mục tiêu bài toán

Thu thập dữ liệu hình ảnh để phục vụ cho các mục dịch khác nhau như:

1. Nghiên cứu và phân tích dữ liệu.

2. Phát triển các ứng dụng liên quan đến hình ảnh.

3. Tạo và cập nhật cơ sở dữ liệu.

4. Phân tích nội dung và nhận diện hình ảnh.

5. Xây dựng và phát triển các mô hình học máy và trí tuệ nhân tạo liên quan đến hình ảnh.

## 3. Các bước thực hiện

1. Truyền key để tìm ảnh.

2. Gán nhãn một phần dữ liệu ảnh để huấn luyện mô hình.

3. Đưa ảnh chưa gán nhãn vào mô hình để dự đoán.

4. Lưu ảnh về.