HỌ VÀ TÊN SINH VIÊN: TRỊNH VINH TUẤN ĐẠT

LỚP: D21CQCN07-B

BỘ THÔNG TIN VÀ TRUYỀN THÔNG

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÁO CÁO
THỰC TẬP CƠ SỞ

Đề tài: Thu thập dữ liệu cá nhân trên nền tảng Facebook

Người hướng dẫn :	**ThS. Đinh Xuân Trường**
Sinh viên thực hiện :	**Trịnh Vinh Tuấn Đạt**
Mã sinh viên :	**B21DCCN031**
Lớp :	**D21CQCN07-B**
Hệ :	**Đại học chính quy**

HÀ NỘI - 2024

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Điểm: **(Bằng chữ:)**

Hà Nội, ngày tháng năm 20...

Giảng viên hướng dẫn

LỜI CẢM ƠN

Em muốn gửi lời cảm ơn chân thành nhất tới thầy cô, bạn bè và những người đã cùng theo dõi, ủng hộ em trong quá trình thực hiện và hoàn thành bài tập lớn.

Đầu tiên và quan trọng nhất, em xin gửi lời cảm ơn sâu sắc đến thầy Đinh Xuân Trường. Cảm ơn thầy đã dành thời gian và tâm huyết để hướng dẫn chúng em những bài giảng hay, kiến thức bổ ích. Nhờ sự hỗ trợ vô cùng quý báu của thầy trong suốt học kì vừa qua, đó thực sự là một nguồn động viên vô cùng to lớn giúp cá nhân em vượt qua những khó khăn trong quá trình triển khai, tìm hiểu và hoàn thành bài tập lớn. Em thực sự rất trân trọng sự tận tâm của thầy đã giúp cá nhân em có một trải nghiệm học tập đáng nhớ.

Em cũng muốn gửi lời cảm ơn tới bạn bè xung quanh, những người đã tạo nên một môi trường học tập tích cực, theo dõi và ủng hộ cá nhân em trong quá trình hoàn thành bài tập lớn.

Hà Nội, ngày 12 tháng 6 năm 2024

Sinh viên

Trịnh Vinh Tuấn Đạt

MỤC LỤC

DANH MỤC HÌNH VẼ

PHẦN MỞ ĐẦU

Mục tiêu và định hướng cá nhân về quá trình thực tập cơ sở: Củng cố kiến thức cơ sở, kỹ năng làm việc cá nhân và làm việc nhóm trước khi bước vào năm học chuyên ngành, tạo tiền đề để cho kỳ thực tập tiếp theo và công việc tương lai sau này.

Trình bày phần đặt vấn đề liên quan đến đề tài của TTCS: Trong thời đại mạng xã hội, việc quản lý và hiểu biết về mối quan hệ trong mạng lưới của chúng ta trở nên ngày càng quan trọng. Facebook, với hàng tỷ người dùng trên toàn thế giới, là một trong những nền tảng mạng xã hội lớn nhất và phức tạp nhất hiện nay. Tuy nhiên, với sự phát triển của mạng xã hội, việc quản lý và tương tác với một số lượng lớn bạn bè và người theo dõi trên Facebook trở nên khó khăn và đôi khi là không hiệu quả.

Trình bày phần mục tiêu và hướng giải pháp: Để giải quyết vấn đề này, một công cụ sử dụng kỹ thuật cào dữ liệu từ trang cá nhân trên Facebook đã được phát triển. Bằng cách kết hợp các công nghệ như Scrapy và Selenium, công cụ có thể thu thập danh sách bạn bè, thông tin về số lượng lượt tương tác và bình luận trên các bài đăng.Dữ liệu sẽ được tiến hành phân tích để xác định danh sách bạn bè ít tương tác nhất và trả về kết quả.

Trình bày phần đóng góp của bài tập cuối khóa: Hướng tới nhu cầu đó, đề tài cuối kỳ Thực tập cơ sở có tên "***THU THẬP DỮ LIỆU CÁ NHÂN TRÊN NỀN TẢNG FACEBOOK***".

Nội dung trình bày trong báo cáo gồm 3 chương chính:

- Chương 1: Báo cáo tiến độ từng tuần trong quá trình học tập môn Thực tập cơ sở
- Chương 2: Báo cáo sản phẩm cuối cùng, gồm 3 chương
  - Giới thiệu đề tài
  - Xác định bài toán
  - Quy trình thực hiện
- Chương 3: Kết luận quá trình Thực tập bao gồm ưu nhược điểm, kết quả đạt được từ đó rút ra những điều cần cải thiện trong tương lai.

CHƯƠNG 1. BÁO CÁO TIẾN ĐỘ TỪNG TUẦN 1-11

Tóm tắt chương *Chương 1 tổng hợp các báo cáo trong suốt quá trình học tập và nghiên cứu ở môn Thực tập cơ sở. Chương bao gồm hai giai đoạn:* ***giai đoạn 1*** *(từ tuần 1 đến tuần 5), đây là giai đoạn khởi động, tìm hiểu một số lý thuyết quan trọng liên quan đến quá trình phát triển phần mềm, ứng dụng;* ***giai đoạn 2*** *(từ tuần 6 đến tuần 11), lựa chọn đề tài cuối kỳ và làm việc nhóm theo nhóm được chỉ định để hoàn thiện sản phẩm được đề ra.*

1.1 Tổng quan quá trình học tập trong 11 tuần

Nội dung thực tập trong 4 tháng

Tiến độ thực hiện thực tập thể hiện trong bảng dưới đây:

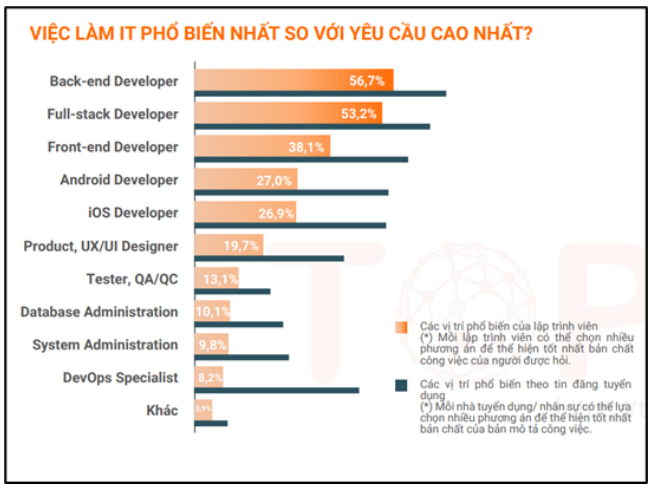
#	Topics	Tuần 1	Tuần 2+3	Tuần 4+5	Tuần 6	Tuần 7	Tuần 8	Tuần 9	Tuần 10+11	...
1	Tìm hiểu về Fullstack Developer									
2	Phân loại phương tiện giao thông với CNN cơ bản									
3	Quy trình phát triển phần mềm Agile									
4	Làm việc với Git									
5	**Lựa chọn đề tài cuối kỳ:Phân loại rác thải**									
6	**Lựa chọn đề tài cuối kì: Hệ thống nhận diện đối tượng đi vào khu vực cấm**									
7	**Tổng quan về scrapy**									
8	**Tổng quan về Splash-Selenium**									
9	**Lựa chọn đề tài:Crawl dữ liệu cá nhân trên nền tảng facebook**									

Hình 1.1: Tiến độ thực hiện thực tập

1.1.1 Tuần 1 (04/03 - 10/03)

Chủ đề tìm hiểu tuần 1: Vẽ bức tranh tổng quát: tìm hiểu về Fullstack Developer

Lý do lựa chọn chủ đề: xu hướng thị trường hiện nay ưa chuộng vị trí Fullstack Developer cũng như để phục vụ cho đề tài phát triển Website trắc nghiệm cuối kỳ. Thống kê sau cho thấy nhu cầu vị trí Fullstack Developer có thứ hạng cao:



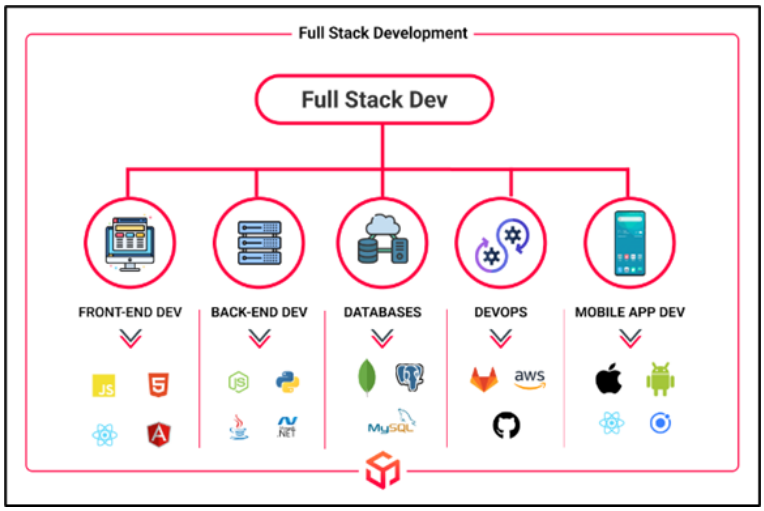
Hình 1.2: Thống kê về việc làm IT so với yêu cầu 2023

Mục tiêu tuần 1: Trình bày tổng quan về vị trí Fullstack Developer, tiềm năng, cơ hội cũng như thách thức và những yêu cầu đối với vị trí này.

Báo cáo kết quả tuần 1:

a. Xu hướng hiện tại

Trước hết, Full Stack là một thuật ngữ tiếng Anh, được ghép từ hai từ "full"(toàn bộ) và "stack"(nhóm). Fullstack Developer là gì? Một Fullstack Developer là người có thể làm việc trên cả front-end và back-end của một ứng dụng. Front-end nói chung là phần mà người dùng có thể thấy được và tương tác được, còn back-end là phần ứng dụng xử lý logic, tương tác cơ sở dữ liệu, chứng thực người dùng, cấu hình máy chủ,...



Hình 1.3: Minh họa về các kiến thức cần phải có của một Fullstack Developer

Full-stack developer tạo các trang web cũng như ứng dụng cho nhiều nền tảng khác nhau. Mô tả công việc của full-stack developer có thể bao gồm những điều sau:

- Phát triển và duy trì các dịch vụ và giao diện web.
- Tạo ra các tính năng mới hoặc các giao diện lập trình ứng dụng (APIs) để mở rộng hoặc cải thiện sản phẩm hiện có.

- Tạo và duy trì máy chủ và cơ sở dữ liệu cho phần Backend của phần mềm.
- Thực hiện kiểm tra, khắc phục sự cố phần mềm và sửa lỗi.
- Phối hợp với các bộ phận khác trong các dự án.

b. Các dự đoán tương lai

Trong những năm gần đây và cả trong tương lai, xu hướng về Trí tuệ nhân tạo và Dữ liệu lớn sẽ bùng nổ và phát triển nhanh chóng. Và việc ứng dụng AI vào đời sống hàng ngày sẽ trở thành một vấn đề ưu tiên hàng đầu đối với lĩnh vực Công nghệ thông tin. Để làm được điều đó, vai trò của software engineer nói chung và fullstack developer nói riêng ngày càng quan trọng hơn, từ đó vạch ra những hướng đi mới, đồng thời tiếp tục cải tiến, nâng cấp công nghệ nền tảng.

c. Cơ hội

Theo như các bản báo cáo, thống kê tổng quan về ngành CNTT nói chung và vị trí Fullstack Developer nói riêng, các số liệu đều cho thấy ngành này đang có tốc độ phát triển rất ấn tượng, thị trường liên tục được mở rộng khắp từ web, phần mềm, di động, và đặc biệt là lĩnh vực học máy.

Hình 1.4: Những vị trí có tốc độ phát triển nhanh nhất trong lĩnh vực IT từ 2023-2025

Vị trí Fullstack Developer thường được đánh giá cao trong ngành CNTT, và do đó thường đi kèm với mức lương hấp dẫn. Với kỹ năng và kinh nghiệm phù hợp, Fullstack Developers có thể đạt được mức lương cao và có tiềm năng tăng lương nhanh chóng.

d. Thách thức

Gần đây, với sự phát triển vượt bậc của Trí tuệ nhân tạo (AI), rất nhiều công việc đã bị thay thế và có nguy cơ bị thay thế bởi công nghệ này. AI hiện nay không chỉ đơn thuần là tự động hóa, mà nó còn có thể học được, cập nhật thông tin nhanh hơn con người rất nhiều mà đối với lĩnh vực CNTT, thì dữ liệu và thông tin chính là sức mạnh. Vì vậy, đối với Fullstack Developer thì AI vừa là công cụ phục vụ cho công việc cũng vừa là mối đe dọa đến nhu cầu việc làm của vị trí này. Có thể AI hiện nay chưa có khả năng nhìn ra được cấu trúc tổng thể của dự án, chưa thể bảo hành bảo trì những hệ thống cũ như một kỹ sư thực thụ nhưng nó đã có khả năng tự xây dựng được những hệ thống đơn giản, lập trình mã nguồn rất nhanh và điều kinh khủng nhất là sự phát

triển của nó sẽ càng lúc càng nhanh và không bao giờ dừng lại. Vì thế, AI là công cụ cho Fullstack Developer hay nó là thứ thay thế Fullstack Developer thì điều đó phụ thuộc vào chính cách nhìn nhận và năng lực của mỗi người đang theo đuổi vị trí này.

e. Nền tảng kiến thức cần có

Đầu tiên, Lập trình viên full-stack là chuyên gia phần mềm mà thông thạo cả hai mảng front-end (phía người dùng) và back-end (phía server). Lập trình full-stack đòi hỏi người đó phải quen thuộc với mọi khâu trong quy trình chế tạo, gia công phần mềm. Họ biết mỗi "stack" sẽ vận hành ra sao và quan trọng nhất là họ có thể điều khiển các thành phần trong back-end.Trở thành một lập trình viên full-stack cần rất nhiều các kỹ năng khác nhau. Cơ bản có thể liệt kê ra bao gồm front-end, back-end,..

- Ngôn ngữ frontend và framework
- Back-end technologies and framework
- Các công cụ khác:
  - Database
  - Version control
  - Web hosting platforms

1.1.2 Tuần 2+3 (11/03 - 24/03)

Chủ đề tìm hiểu tuần 2+3: Nhận diện hình ảnh phương tiện giao thông với CNN
Lý do lựa chọn đề tài: nhóm chúng em muốn thử tìm hiểu ứng dụng của CNN vào trong nhận diện hình ảnh, với dự tính có thể sẽ ứng dụng vào đề tài ứng dụng cuối kỳ của nhóm.

Mục tiêu tuần 2+3: Sử dụng CNN xử lý một bài toán đơn giản về nhận diện hình ảnh, ở đây bài toán chúng em lựa chọn là nhận diện phương tiện giao thông.

Báo cáo kết quả tuần 2+3:

a. Giới thiệu bài toán, lý thuyết về CNN

Giới thiệu bài toán: Xây dựng ứng dụng nhận diện các phương tiện giao thông dựa trên Computer Vision bằng việc sử dụng hình ảnh kỹ thuật số từ các máy ảnh và những đoạn phim cũng như các mô hình học sâu (deep learning).

- Input: Hình ảnh các phương tiện giao thông (gồm 3 loại: xe đạp, xe máy, ô tô).
- Output: Xác định tên loại phương tiện.

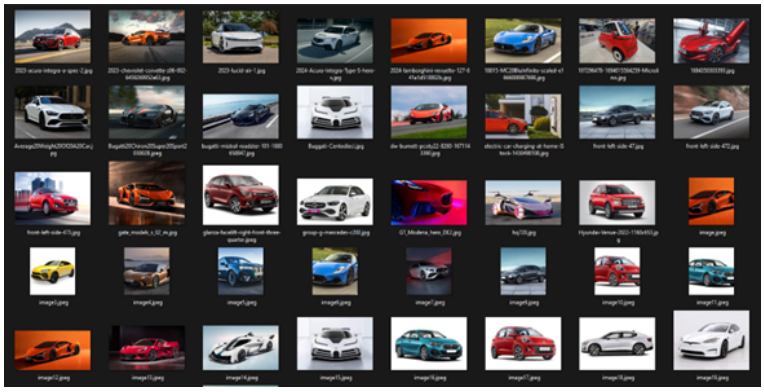
Thuật toán CNN: Mạng nơ-ron tích chập hay còn gọi là ConvNet/CNN, đây là một trong những mô hình của học sâu (Deep learning) vô cùng tiên tiến.

Đặc trưng (Feature): CNN phân loại hình ảnh bằng cách lấy một hình ảnh đầu vào, xử lý và phân loại nó theo các hạng mục nhất định. Khi nhập dữ liệu hình ảnh vào, máy tính sẽ chia nhỏ hình ảnh đó ra thành một mảng hình ảnh nhỏ hơn, mỗi mảng đó được gọi là feature.

Tích chập (Convolutional): để thuật toán CNN nhận biết được hình ảnh ở vị trí nào, các feature khớp với nhau ở đâu thì tích chập sẽ thử chúng với tất cả các vị trí khác nhau và tạo thành một bộ lọc gọi là filter. Quá trình này được thực hiện thông qua phép toán nơ-ron tích chập.

b. Xử lý dữ liệu và xây dựng mô hình nhận diện

Thu thập dữ liệu: Các thành viên trong nhóm đã tập trung tìm kiếm và tải xuống từ Google photos các hình ảnh của 3 loại phương tiện giao thông: ô tô, xe máy, xe đạp. Nhóm sử dụng các hình ảnh có màu, sau đó chia các hình ảnh này thành các nhóm và thực hiện tải thư mục lên Google drive. Hình ảnh mà nhóm thu thập có dạng như sau:



Hình 1.5: Hình ảnh ô tô con nhóm đã thu thập

Nhóm đã chia dữ liệu thu thập được thành 3 tệp: Ô Tô, Xe Đạp, Xe Máy. Tổng số lượng là 438 ảnh, tỷ lệ Train:Validation là 80/20 trong đó:

- Tệp Train gồm 300 ảnh: sử dụng để huấn luyện mô hình học sâu.
- Tệp Validation gồm 75 ảnh: sử dụng để đánh giá mô hình sau mỗi lần huấn luyện.
- Ngoài ra còn một lớp Test sử dụng để kiểm thử kết quả huấn luyện có 63 ảnh.

Xây dựng mô hình: sử dụng Python và môi trường Google Notebook

- Khai báo các thư viện sử dụng

```
import tensorflow as tf
from tensorflow import keras
import matplotlib.pyplot as plt
import numpy as np
from google.colab import drive
import os
```

Hình 1.6: Các thư viện python sử dụng

- Kết nối với Google Drive để đọc và lưu dữ liệu ảnh

```
[3] drive.mount("/content/drive")
    train_image_files_path = "/content/drive/MyDrive/Date/Train/"
    valid_image_files_path = "/content/drive/MyDrive/Date/Validation/"
```

Hình 1.7: Kết nối tới thư mục Train và Validation trong drive lưu dữ liệu ảnh

- Gán nhãn dữ liệu: Đối với một mô hình đưa ra quyết định và thực hiện hành động nó phải được đào tạo để hiểu thông tin cụ thể. Phân loại ảnh là bài toán học có giám sát, do đó dữ liệu huấn luyện và kiểm

định phải được gán nhãn. Gán nhãn dữ liệu là quá trình gắn tường ý nghĩa cho các loại dữ liệu kỹ thuật số khác nhau như tệp âm thanh, văn bản, hình ảnh, video, . . . Tên và thứ tự các nhãn sẽ tương ứng với tên và thứ tự các thư mục ảnh huấn luyện và kiểm định.

```
label = ['O to', 'Xe Dap', 'Xe May']
```

Hình 1.8: Đặt tên nhãn cho mô hình phân loại

- Tiền xử lý dữ liệu hình ảnh với ImageDataGenerator và đọc dữ liệu vào mô hình

```
from tensorflow.keras.preprocessing.image import ImageDataGenerator
train_data_gen = ImageDataGenerator(
    rescale=1/255,
)
validation_data_gen = ImageDataGenerator(
    rescale=1/255,
)
test_data_gen = ImageDataGenerator(rescale=1/255)
train_generator = train_data_gen.flow_from_directory(
    train_image_files_path,
    target_size=(224, 224),
    batch_size=32,
    class_mode="categorical",
    shuffle=True,
    seed=42
)
validation_generator = validation_data_gen.flow_from_directory(
    valid_image_files_path,
    target_size=(224, 224),
    batch_size=32,
    class_mode="categorical",
    shuffle=True,
    seed=42
)
test_generator = test_data_gen.flow_from_directory(
    test_image_files_path,
    target_size=(224, 224),
    class_mode="categorical"
)
```

Hình 1.9: Chuẩn hóa dữ liệu ảnh đầu vào cho mô hình

Kết quả chạy:

```
Found 300 images belonging to 3 classes.
Found 75 images belonging to 3 classes.
Found 63 images belonging to 3 classes.
```

Hình 1.10: Kết quả chạy đoạn mã đọc dữ liệu

- Dựng mô hình CNN

```
from keras.models import Sequential
from keras.layers import Dense, Dropout, Conv2D, MaxPooling2D, Flatten, BatchNormalization
model = tf.keras.models.Sequential()

# lớp CNN1
model.add(Conv2D(32, (3,3), activation='relu', input_shape=(224,224,3)))
model.add(Dropout(0.25))
model.add(MaxPooling2D(2,2))

# lớp CNN2
model.add(Conv2D(64, (3,3), activation='relu'))
model.add(Dropout(0.25))
model.add(MaxPooling2D(2,2))

# lớp CNN3
model.add(Conv2D(128,(5,5), activation='relu'))
model.add(Dropout(0.25))
model.add(MaxPooling2D(2,2))

# # lớp CNN4
# model.add(Conv2D(256,(3,3), activation='relu'))
# model.add(MaxPooling2D(2,2))

model.add(Flatten())

# lớp ẩn
model.add(Dense(512, activation=tf.nn.relu))
# lớp output
model.add(Dense(5, activation=tf.nn.softmax))
```

Hình 1.11: Mô hình nhận diện ảnh bằng CNN gồm 3 lớp CNN

Mô hình gồm 5 lớp CNN1 => CNN2 => CNN3 => Fully connected layer => Output

- Huấn luyện mô hình

```
from tensorflow.keras.optimizers import Adamax
model.compile(optimizer=Adamax(lr=0.001),
              loss='categorical_crossentropy',
              metrics=['acc'])
```

Hình 1.12: Tham số huấn luyện mô hình

```
from keras.models import load_model
EPOCHS=80
history=model.fit(
    train_generator,
    steps_per_epoch=2,
    epochs=EPOCHS,
    verbose=1,
    validation_data=validation_generator,
    validation_steps=2)
```

Hình 1.13: Huấn luyện mô hình với luồng dữ liệu train_generator

Kết quả khi chạy đoạn mã trên để train mô hình:

```
Epoch 1/80
/usr/local/lib/python3.10/dist-packages/PIL/Image.py:996: UserWarning: Palette images with Transparency expressed in by
  warnings.warn(
2/2 [==============================] - 9s 3s/step - loss: 45.0819 - acc: 0.3125 - val_loss: 5.1146 - val_acc: 0.2812
Epoch 2/80
2/2 [==============================] - 3s 3s/step - loss: 8.8596 - acc: 0.2656 - val_loss: 1.5690 - val_acc: 0.3594
Epoch 3/80
2/2 [==============================] - 2s 2s/step - loss: 2.7959 - acc: 0.2656 - val_loss: 1.2085 - val_acc: 0.3438
Epoch 4/80
2/2 [==============================] - 2s 2s/step - loss: 1.3944 - acc: 0.3281 - val_loss: 1.0957 - val_acc: 0.3906
Epoch 5/80
2/2 [==============================] - 4s 3s/step - loss: 1.1368 - acc: 0.2727 - val_loss: 1.0984 - val_acc: 0.3594
Epoch 6/80
2/2 [==============================] - 3s 3s/step - loss: 1.0924 - acc: 0.3182 - val_loss: 1.0964 - val_acc: 0.4062
Epoch 7/80
2/2 [==============================] - 2s 2s/step - loss: 1.0930 - acc: 0.4318 - val_loss: 1.0952 - val_acc: 0.5469
Epoch 8/80
2/2 [==============================] - 2s 2s/step - loss: 1.0892 - acc: 0.4219 - val_loss: 1.0925 - val_acc: 0.3750
Epoch 9/80
2/2 [==============================] - 2s 2s/step - loss: 1.0822 - acc: 0.3125 - val_loss: 1.0882 - val_acc: 0.3750
Epoch 10/80
2/2 [==============================] - 3s 3s/step - loss: 1.0742 - acc: 0.3636 - val_loss: 1.0847 - val_acc: 0.3906
Epoch 11/80
2/2 [==============================] - 3s 3s/step - loss: 1.0582 - acc: 0.3906 - val_loss: 1.0767 - val_acc: 0.5000
Epoch 12/80
2/2 [==============================] - 3s 3s/step - loss: 1.0434 - acc: 0.4844 - val_loss: 1.0721 - val_acc: 0.4375
Epoch 13/80
2/2 [==============================] - 3s 3s/step - loss: 0.9743 - acc: 0.5455 - val_loss: 1.0650 - val_acc: 0.5156
Epoch 14/80
2/2 [==============================] - 3s 3s/step - loss: 0.9873 - acc: 0.5156 - val_loss: 1.0369 - val_acc: 0.6250
```

Hình 1.14: Kết quả huấn luyện mô hình

Đánh giá: sau 80 epoch, giá trị loss giảm xuống thấp 0.0078, giá trị acc cao 1.0000 cho thấy mô hình hoạt động tốt trên tập Train. Tuy nhiên giá trị val_loss 0.7861 và val_acc 0.7188 không quá khả quan cho thấy mô hình chưa đạt kết quả tốt trên tập Validation, lý do có thể vì bộ dữ liệu chưa đủ lớn và các tham số chưa được tối ưu.

Sau khi train mô hình xong, thực hiện lưu mô hình:

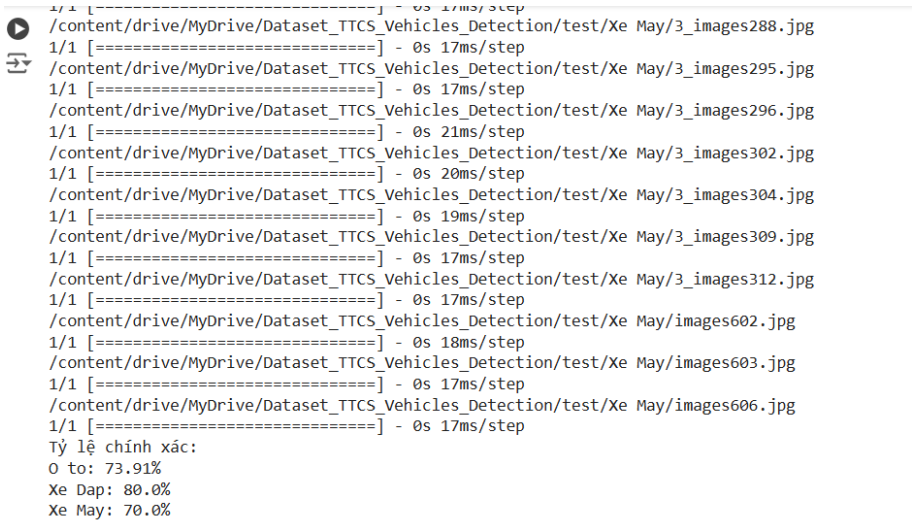
```
# Lưu model
from keras.models import load_model
path='/content/drive/MyDrive/Dataset_TTCS_Vehicles_Detection/mdel6.h5o'
model.save(path)
```

Hình 1.15: Lưu mô hình vào file đuôi .h5

- Sử dụng mô hình để test ảnh: Sử dụng 60 ảnh từ tệp Test đưa vào mô hình để chạy kiểm tra thử độ chính xác khi sử dụng thực tế:

```
[ ] from google.colab import files
    from keras.preprocessing import image
    %matplotlib inline
    import matplotlib.pyplot as plt
    import matplotlib.image as mpimg
    import numpy as np
    from keras.models import load_model
    model = load_model('/content/drive/MyDrive/Dataset_TTCS_Vehicles_Detection/model6.h5')
    soluong_model = 3
    correct_predict = np.zeros(soluong_model)
    total = np.zeros(soluong_model)
    for i, path in enumerate(test_generator.filepaths):
      # i là số thứ tự ảnh trong thư mục Test
      number_label = test_generator.labels[i]
      # number_label là số thứ tự của nhãn
      total[number_label] += 1
      img=image.load_img(path, target_size=(224,224))
      print(path)
      img_arr = image.img_to_array(img)
      img_arr = np.expand_dims(img_arr,axis=0)
      images=np.vstack([img_arr])
      y_predict=model.predict(images,batch_size=10)
      number_label_predict = np.argmax(y_predict)
      if number_label_predict == number_label:
        correct_predict[number_label_predict] += 1
    correct_accuracies = correct_predict / total
    print("Tỷ lệ chính xác:")
    for i in range(soluong_model):
      print(label[i] + ': ' + str(round(correct_accuracies[i]*100, 2)) + '%')
```

Hình 1.16: Đoạn mã đưa ảnh vào mô hình để kiểm tra thử



Hình 1.17: Kết quả nhận được với 60 ảnh từ tệp Test

c. Kết luận

Mô hình nhận diện phương tiện giao thông gồm 3 lớp CNN cấu tạo khác nhau, huấn luyện sau khoảng 80 Epochs với bộ dữ liệu hình ảnh tải xuống từ Google (hơn 400 ảnh), cho kết quả không quá khả quan, tỷ lệ chính xác của mô hình rất cao (75%) khi thực hiện kiểm tra thử nghiệm trên bộ Test (60 ảnh). Thời gian huấn luyện mô hình tương đối nhanh.

Các tham số trong đoạn mã nguồn của mô hình được lựa chọn bằng cách thực nghiệm, huấn luyện mô hình liên tục để lựa chọn ra giá trị cho kết quả khả quan nhất.

Kết quả từ việc xây dựng mô hình CNN trên cho chúng em nhiều kinh nghiệm quý giá giúp ích việc xây dựng mô hình học máy sau này.

1.1.3 Tuần 4+5 (25/03 - 07/04)

Chủ đề tìm hiểu tuần 4+5: Gồm 2 chủ đề sau:

- Chủ đề 1: Quy trình phát triển phần mềm Agile
- Chủ đề 2: Công cụ phát triển phần mềm và quản lý mã nguồn. Làm việc với Git, Github

Lý do lựa chọn các chủ đề trên: nhóm chúng em muốn tìm hiểu quá trình làm việc nhóm khi xây dựng một phần mềm, ứng dụng nhằm phục vụ cho đề tài cuối kỳ được thực hiện một cách chuẩn chỉ và trơn tru hơn.

Mục tiêu tuần 4+5: Hiểu cơ bản về quy trình Agile và các bước sử dụng phổ biến với Git, Github khi ứng dụng vào làm việc nhóm.

Báo cáo kết quả tuần 4+5:

a. Quy trình phát triển phần mềm Agile

Sử dụng phần mềm Jira để quản lý quy trình phát triển phần mềm dựa trên phương pháp Scrum:

- Tạo 1 Scrum project với Jira
- Tạo User stories và tasks trong backlog
- Tạo một Sprint

- Tổ chức Sprint Planning Meeting
- Bắt đầu một Sprint với Jira
- Xem biểu đồ Burndown
- Hoàn thành Sprint

b. Công cụ phát triển phần mềm và quản lý mã nguồn, làm việc với Git, Github

Công cụ phát triển phần mềm và quản lý mã nguồn

Trong ngành công nghiệp phần mềm, việc quản lý mã nguồn là một phần không thể thiếu trong quá trình phát triển dự án.

Hệ thống quản lý mã nguồn (Source Code Management System - SCM):

Khái niệm: SCM là một hệ thống hoặc công cụ được sử dụng để quản lý và theo dõi mã nguồn trong quá trình phát triển phần mềm. Nó cho phép các nhà phát triển làm việc song song trên cùng một dự án và theo dõi lịch sử thay đổi.

Vai trò: SCM giúp quản lý phiên bản, nhánh, hợp nhất và theo dõi lịch sử thay đổi của mã nguồn. Nó cung cấp một cơ sở để hợp tác, phát triển và duy trì mã nguồn dễ dàng.

Git và GitHub đã trở thành công cụ quản lý mã nguồn phổ biến và mạnh mẽ, giúp các chuyên gia phát triển, quản lý và theo dõi mã nguồn một cách hiệu quả.

Git

Có bốn khu vực khác nhau trong vòng đời của mã nguồn trong Git:

- Thư mục làm việc (Working Directory)
- Vùng đợi commit (Staging area)
- Kho lưu trữ (.git thư mục)
- Kho lưu trữ từ xa (Remote)

Một số lệnh Git cơ bản:

- Tạo Repository ở local: git init
- Xem trạng thái của repository: git status
- Thêm vào staging: git add [tên file] hoặc git add .
- Tạo commit: git commit -m “Thông điệp commit”.

Github

Tạo một remote reposity: Public nếu công khai hoặc Private không công khai.

Gắn url remote vào git: git remote add origin [url_remote].

Đổi tên nhánh hiện tại thành main: git branch -M main.

Đẩy các thay đổi lên nhánh main remote: git push -u origin main.

1.1.4 Tuần 6 (08/04 - 14/04)

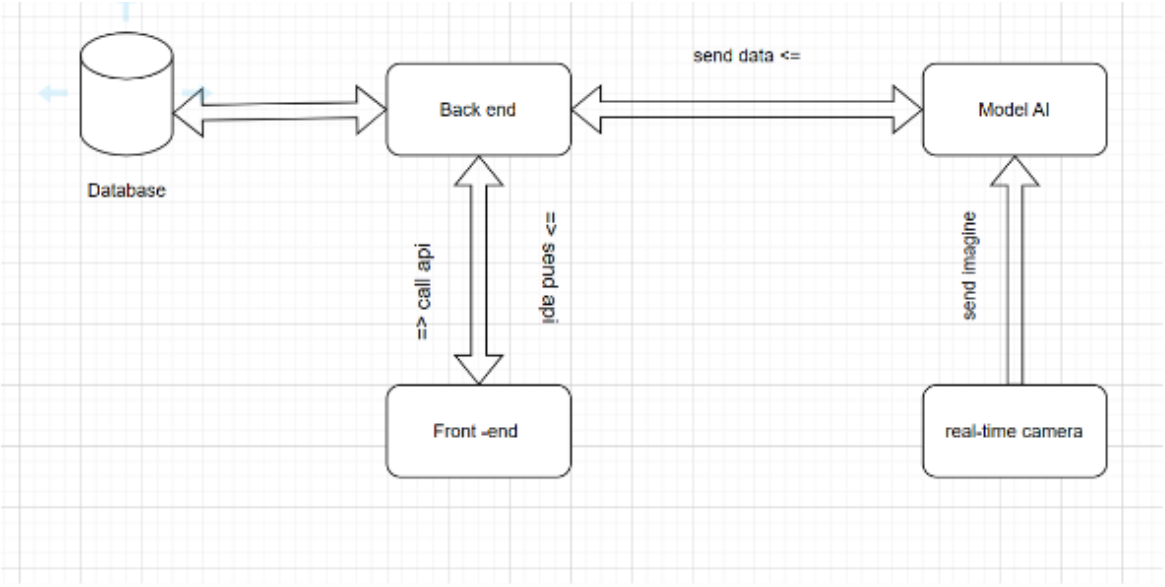
Chủ đề tuần 6: Hệ thống nhận diện rác thải kết hợp với web quản lý rác thải Rác thải vẫn đã và đang là một vấn đề nhức nhối đối với đời sống con người. Với sự phát triển nhanh của xã hội, rác thải được sinh ra nhiều hơn

=> phân loại, tái chế và quản lý là một điều cấp bách và cần thiết để giảm thiểu rác thải

Mục tiêu tuần 6: Trình bày khái quát về Hệ thống nhận diện rác thải kết hợp với web quản lý rác thải

Nội dung:

1. Nền tảng triển khai:
   - Đối với mô hình nhận diện, sử dụng yolo v8 để train model.
   - Đối với hệ thống web, sẽ sử dụng flask để triển khai web.
   - Cơ sở dữ liệu: Mysql.
2. Luồng hoạt động:



Hình 1.18: Luồng hoạt động

3. Kết luận:
   Nhìn chung mô hình sẽ giúp cải thiện số lượng rác thải, giúp tăng cường việc bảo vệ môi trường và chung tay hướng đến một thế giới xanh

1.1.5 Tuần 7 (15/04 - 21/04)

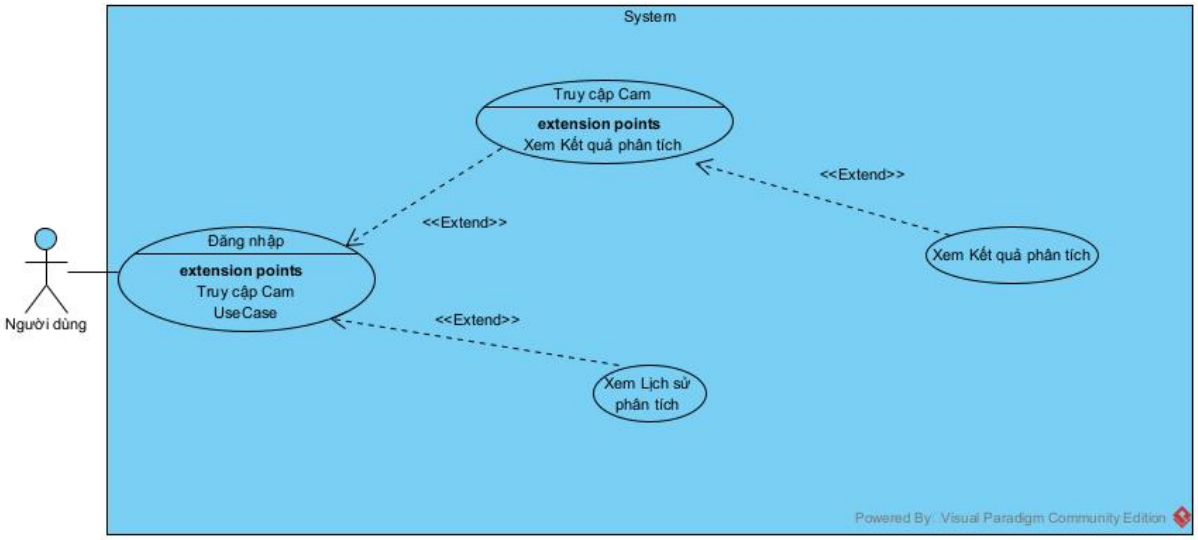
Chủ đề : Ứng dụng kiểm soát và phát hiện đối tượng đi vào khu vực cấm Để cải thiện an ninh cho các khu vực quan trọng như: nhà máy, khu vực cấm ,nhà ở ,... Việc kiểm soát và phát hiện đối tượng là một phần quan trọng của việc đảm bảo an ninh.

Mục tiêu tuần 7: Trình bày khái quát về ứng dụng kiểm soát và phát hiện đối tượng đi vào khu vực cấm

Nội dung:

1. Đối tượng:
   - Phát hiện đối tượng là người đi bộ, nhận diện các hành vi bất thường.
   - Mô hình sẽ phát hiện đối tượng người đi bộ đi vào vùng cấm được vẽ trên khung hình video/webcam và đưa ra tin nhắn cảnh báo.

2. Chức năng:
    - Phát hiện đối tượng trong tầm nhìn của camera.
    - Lưu trữ lịch sử các lần đối tượng xâm nhập trái phép.
3. Yêu cầu phi chức năng:
    - Cung cấp các cơ chế cảnh báo nhanh chóng và chính xác.
    - Tương thích và dễ dàng triển khai.
    - Độ Tin Cậy và Khả Năng Điều Khiển Từ Xa.
4. Công nghệ:
    - Mô hình học máy: Sử dụng YOLO.
    - Ứng dụng phần mềm: Web.
5. Use case:



Hình 1.19: Use Case tổng quan

6. Kết luận:

   Mô hình hướng tới mong muốn đảm bảo an ninh, giúp bảo vệ người dùng khỏi những môi hiểm họa tiềm ẩn.

1.1.6 Tuần 8 (22/04 - 28/04)

Chủ đề tìm hiểu tuần 8: Tìm hiểu về Scrapy

Mục tiêu tuần 8: Trình bày khái quát về crawl data, tổng quan, cấu trúc và cách hoạt động của scrapy

Kết quả của tuần 8:

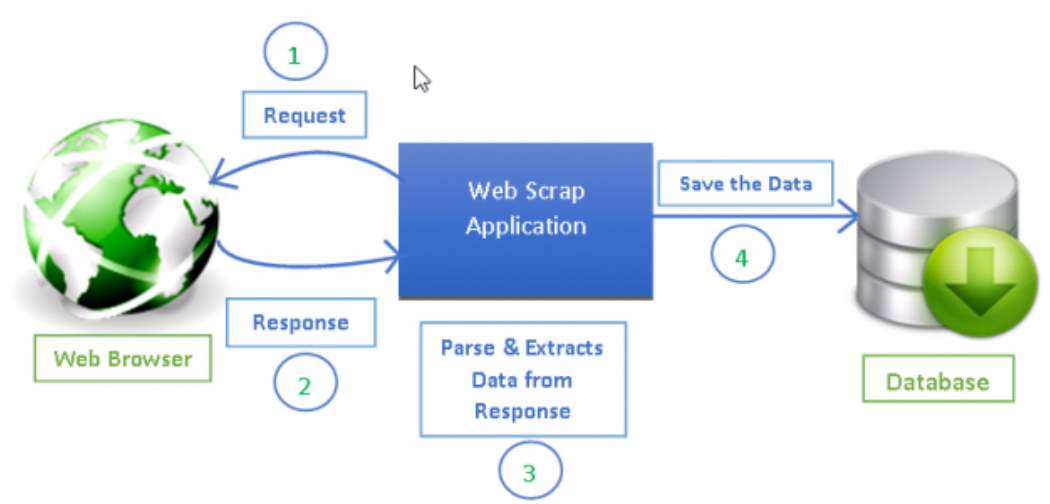
1. Giới thiệu về scrapy

   (a) Giới thiệu

   Scrapy là một framework mã nguồn mở mạnh mẽ dành cho việc thu thập dữ liệu (web scraping) và khai thác dữ liệu (data extraction) từ các trang web. Được phát triển bằng ngôn ngữ lập trình Python,

Scrapy cung cấp một bộ công cụ hoàn chỉnh để xây dựng các spider tự động duyệt và thu thập thông tin từ các trang web theo các quy tắc định trước

(b) Mô hình hoạt động



Hình 1.20: Mô hình hoạt động

+) Scapy Application sẽ gửi request lên webserver

+) Web server sẽ gửi trả lại cho response cho Scrapy application

+) Sau khi nhận được dữ liệu, Scrapy Application sẽ phân tích và trích xuất dữ liệu từ response nhận được

+) Lưu trữ dữ liệu vừa phân tích

(c) Cấu trúc

Cấu trúc của scrapy bao gồm 6 thành phần chính:

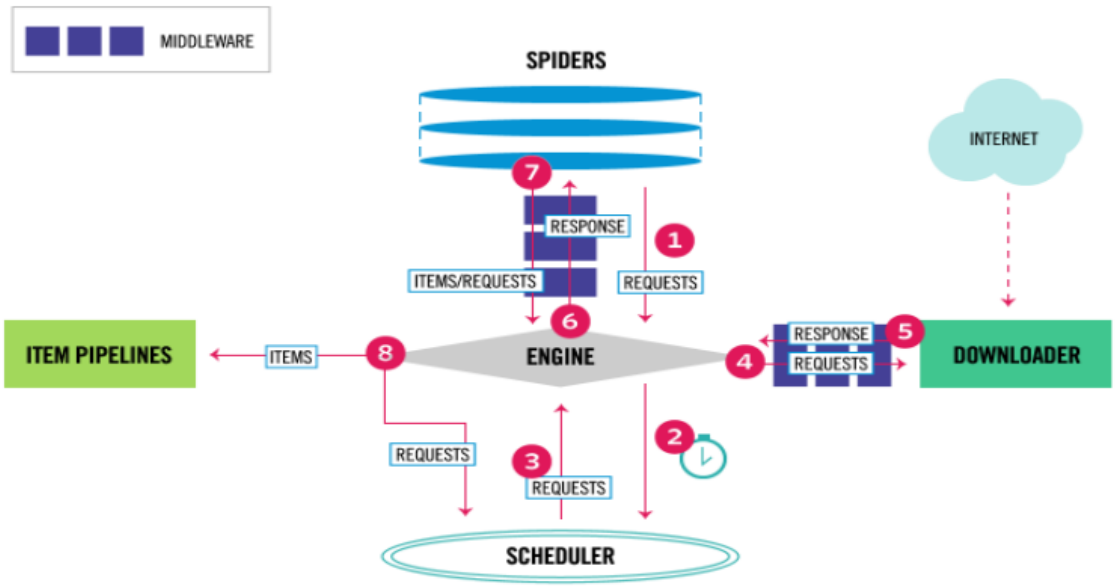
+) Scrapy Engine: có trách nhiệm kiểm soát luồng dữ liệu giữa tất cả các thành phần của hệ thống và kích hoạt các sự kiện khi một số hành động xảy ra

+) Scheduler giống như một hàng đợi (queue), scheduler sắp xếp thứ tự các URL cần download

+) Downloader Thực hiện download trang web và cung cấp cho engine Spiders Spiders là class được viết bởi người dùng, có trách nhiệm bóc tách dữ liệu cần thiết và tạo các url mới để nạp lại cho scheduler qua engine.

+) Item Pipeline Những dữ liệu được bóc tách từ spiders sẽ đưa tới đây, Item pipeline có nhiệm vụ xử lý và lưu vào cơ sở dữ liệu

+) Các Middlewares Là các thành phần nằm giữa Engine với các thành phần khác, đều có mục đích là giúp người dùng có thể tùy biến, mở rộng khả năng xử lý cho các thành phần. VD: sau khi



Hình 1.21: Cấu trúc của scrapy

download xong url, bạn muốn tracking, gửi thông tin ngay lúc đó thì bạn có thể viết phần mở rộng và sửa lại cấu hình để sau khi Dowloader tải xong trang thì sẽ thực hiện việc tracking.

+) Spider middlewares Là thành phần nằm giữa Engine và Spiders, có nhiệm vụ xử lý các response đầu vào của Spiders và đầu ra (item và các url mới).

(d) Ứng dụng

Scrapy có thể được sử dụng trong nhiều ứng dụng khác nhau, bao gồm:

+) Xây dựng công cụ tìm kiếm web: Scrapy có thể được sử dụng để xây dựng các công cụ tìm kiếm web tự động, giúp tạo ra các bản sao của dữ liệu từ các trang web và cung cấp kết quả tìm kiếm nhanh chóng.

+) Monitoring và theo dõi thị trường: Scrapy có thể được sử dụng để theo dõi thị trường bằng cách tự động lấy thông tin từ các trang web cạnh tranh, giúp doanh nghiệp hiểu rõ hơn về động thái của thị trường và cạnh tranh.

+) Phân tích nội dung web: Scrapy có thể được sử dụng để phân tích nội dung web và trích xuất thông tin cụ thể từ các trang web, giúp tổ chức và xử lý dữ liệu web một cách hiệu quả.

+) Xây dựng công cụ đánh giá sản phẩm và dịch vụ: Scrapy có thể được sử dụng để tự động thu thập đánh giá và phản hồi từ các trang web thương mại điện tử hoặc diễn đàn, giúp doanh nghiệp hiểu rõ hơn về cảm nhận của khách hàng về sản phẩm và dịch vụ.

+) Xây dựng dịch vụ thông tin tự động: Scrapy có thể được sử dụng để xây dựng các dịch vụ thông tin tự động, như tổng hợp tin tức từ nhiều nguồn khác nhau trên internet và cung cấp thông tin đa dạng và cập nhật liên tục cho người dùng.

2. XPath và Css Selector

(a) Giới thiệu

Trong quá trình thu thập dữ liệu từ các trang web, việc trích xuất thông tin từ HTML là một bước quan trọng. XPath và CSS Selector là hai kỹ thuật phổ biến được sử dụng để định vị và trích xuất các phần tử trong cây DOM (Document Object Model) của một trang web.

(b) Xpath

XPath (XML Path Language) là một ngôn ngữ truy vấn được sử dụng để chọn các nút từ một tài liệu XML, có thể áp dụng tương tự với HTML vì cấu trúc của HTML là một dạng của XML. Cú pháp cơ bản XPath cung cấp một cú pháp mạnh mẽ và linh hoạt để truy vấn các phần tử.
Một số cú pháp cơ bản bao gồm:

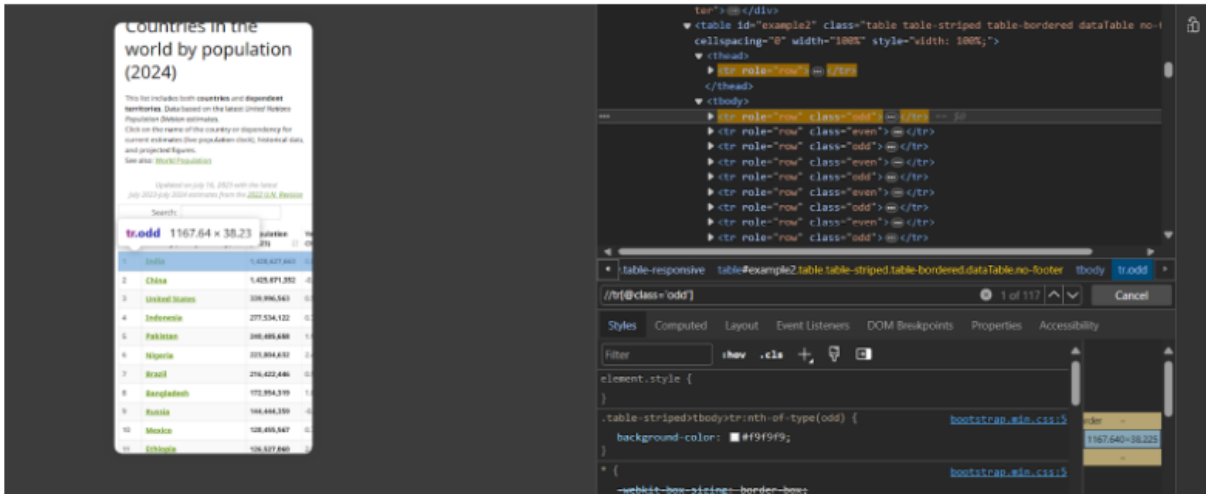
+) //: Chọn các phần tử từ bất kỳ vị trí nào trong tài liệu.

+) /: Chọn các phần tử ngay dưới một phần tử cha.

+) @: Chọn thuộc tính của một phần tử.

+) *: Chọn tất cả các phần tử.

+) []: Định vị phần tử dựa trên các điều kiện.



Hình 1.22: Ví dụ về Xpath

(c) Css Selector

CSS Selector là một cú pháp được sử dụng để chọn các phần tử HTML dựa trên các quy tắc CSS. Nó được sử dụng phổ biến trong cả việc định dạng CSS và truy vấn DOM. Cú pháp cơ bản CSS Selector cung cấp một cú pháp đơn giản và dễ hiểu.
Một số cú pháp cơ bản bao gồm:

+) `tagname: Chọn tất cả các phần tử với tên thẻ cụ thể.`

+) `.classname: Chọn tất cả các phần tử với lớp cụ thể.`

+) `#id: Chọn phần tử với ID cụ thể.`

+) `tagname.classname: Chọn các phần tử với tên thẻ và lớp cụ thể.`

+) `tagname > tagname: Chọn các phần tử con trực tiếp.`

3. Kết luận:

   Nhìn chung, Scrapy là một công cụ hữu ích và hiệu quả cho việc thu thập và xử lý dữ liệu từ web. Với tính linh hoạt, hiệu suất cao và khả năng mở rộng, Scrapy tiếp tục là lựa chọn hàng đầu cho các dự án thu thập dữ liệu trên web của các nhà phát triển

1.1.7 Tuần 9 (29/04 - 05/05)

Chủ đề tìm hiểu tuần 9: Tổng quan về splash - selenium

Mục tiêu tuần 9: Trình bày, nắm rõ được cách thức hoạt động và của splash và selenium

Kết quả của tuần 9:

1. Giới thiệu chung

   Trong quá trình thu thập dữ liệu từ web, có nhiều trang web sử dụng JavaScript để tải và hiển thị nội dung. Các trang web này không thể được xử lý hoàn toàn chỉ bằng các công cụ thu thập dữ liệu HTML tĩnh như Scrapy. Để xử lý các trang web động này, cần sử dụng các công cụ có khả năng render JavaScript như Splash và Selenium.

2. Splash

   (a) Giới thiệu

   Splash là một trình duyệt headless được thiết kế đặc biệt để render nội dung trang web bằng cách thực thi JavaScript. Được phát triển bởi Scrapinghub, Splash cung cấp các API HTTP để tương tác với các trang web động mà các công cụ thu thập dữ liệu truyền thống không thể xử lý trực tiếp do sử dụng JavaScript.

   (b) Kiến trúc

   +) Splash Server: Đây là một ứng dụng web API được viết bằng Python sử dụng Twisted, một thư viện mạng dành cho Python. Splash Server chịu trách nhiệm cho việc thực thi JavaScript và render trang web.

   +) API HTTP của Splash: Splash cung cấp một loạt các API HTTP cho phép tương tác với Splash Server. Các yêu cầu có thể được gửi đến Splash để thực thi JavaScript, render trang web, chụp ảnh màn hình, và truy xuất nội dung đã render của trang.

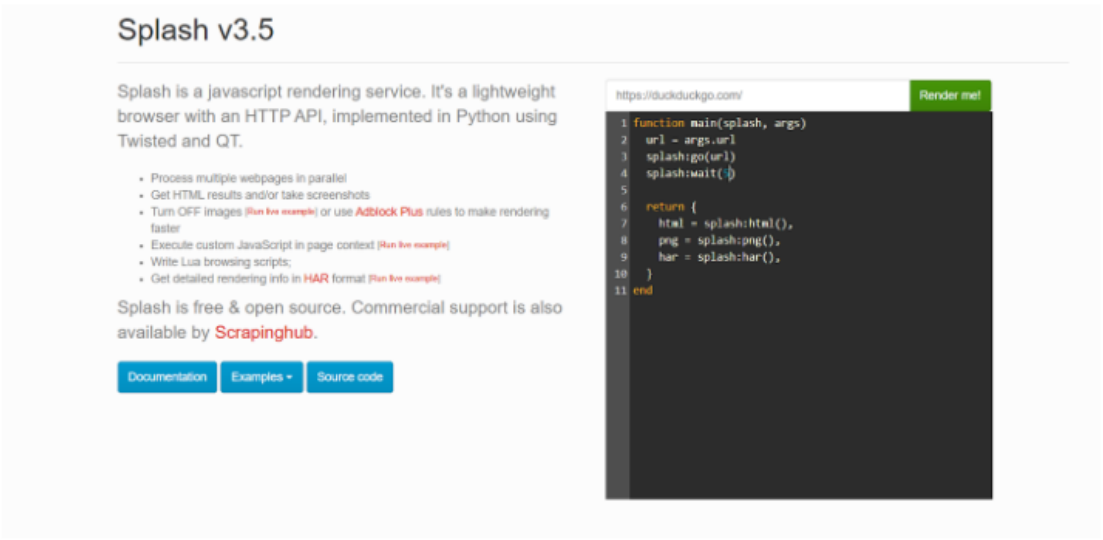
   (c) Cách hoạt động

   +) Yêu cầu từ Client: Client gửi yêu cầu HTTP đến Splash Server.

   +) Splash Render Trang web: Splash Server nhận yêu cầu, thực thi JavaScript và render trang web theo yêu cầu của client.

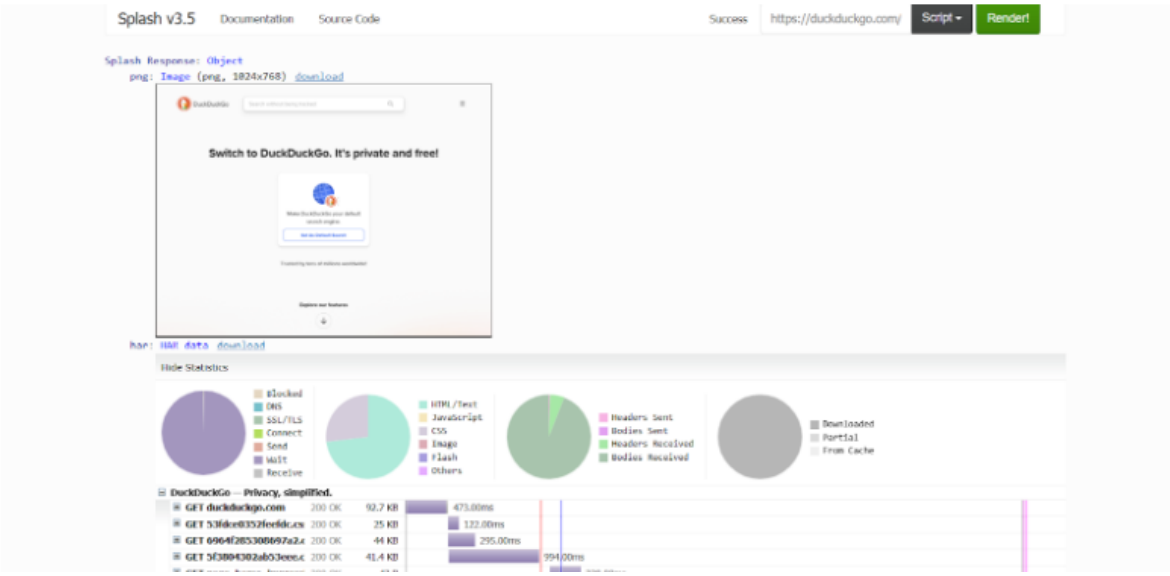
   +) Trả về Kết quả: Sau khi quá trình render hoàn tất, Splash trả về nội dung đã render cho client thông qua API HTTP.

   (d) Sử dụng

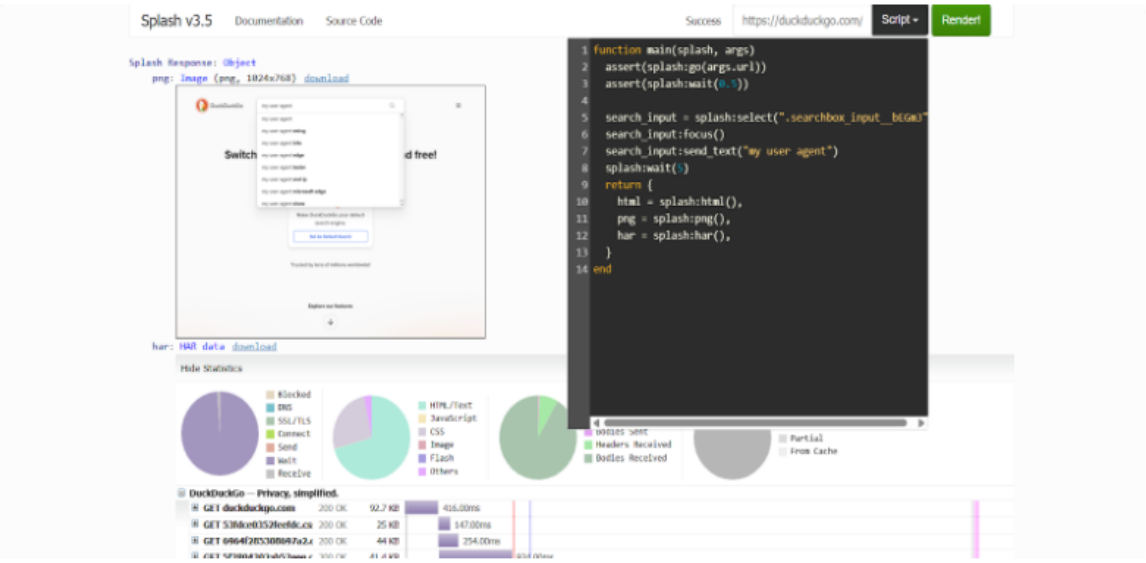


Hình 1.23: Giao diện của splash

Sau khi thực hiện một số câu lệnh đơn giản đối với splash, dữ liệu sẽ được trả về.

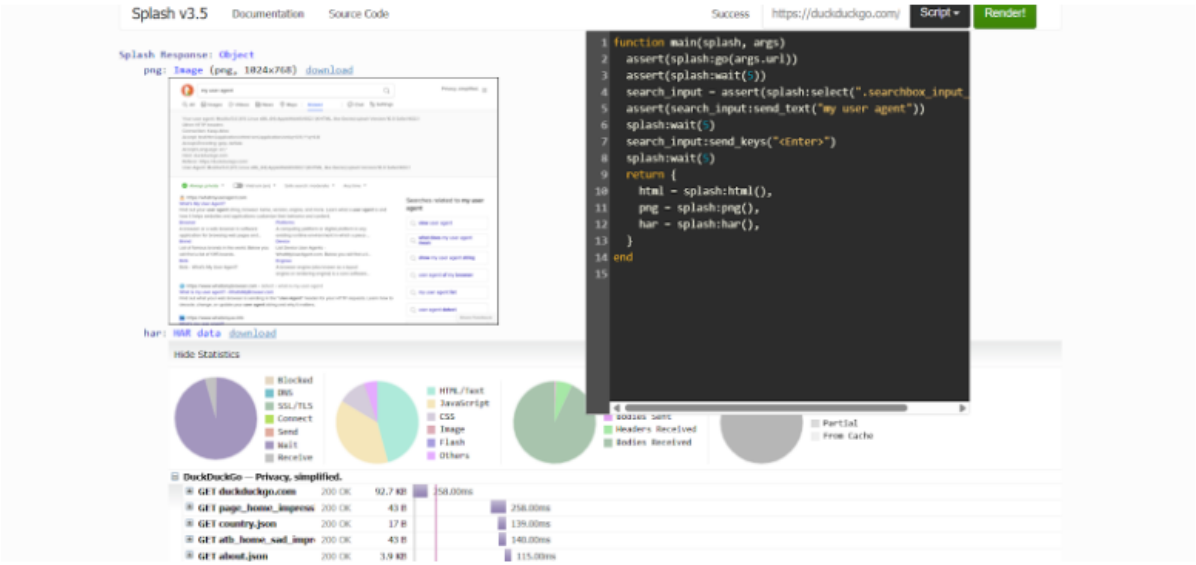


Hình 1.24: Dữ liệu trả về



Hình 1.25: Thao tác cơ bản với splash

Thực hiện xác định ô tìm kiếm , và gửi dòng kí tự "my user agent"đến ô tìm kiếm



Hình 1.26: Dữ liệu trả về

3. Selenium

(a) Giới thiệu

Selenium là một công cụ tự động hóa trình duyệt web phổ biến được sử dụng cho việc kiểm thử tự động, tự động hóa các công việc trên trình duyệt web, và thu thập dữ liệu từ các trang web

Một số ưu điểm khi sử dụng selenium:

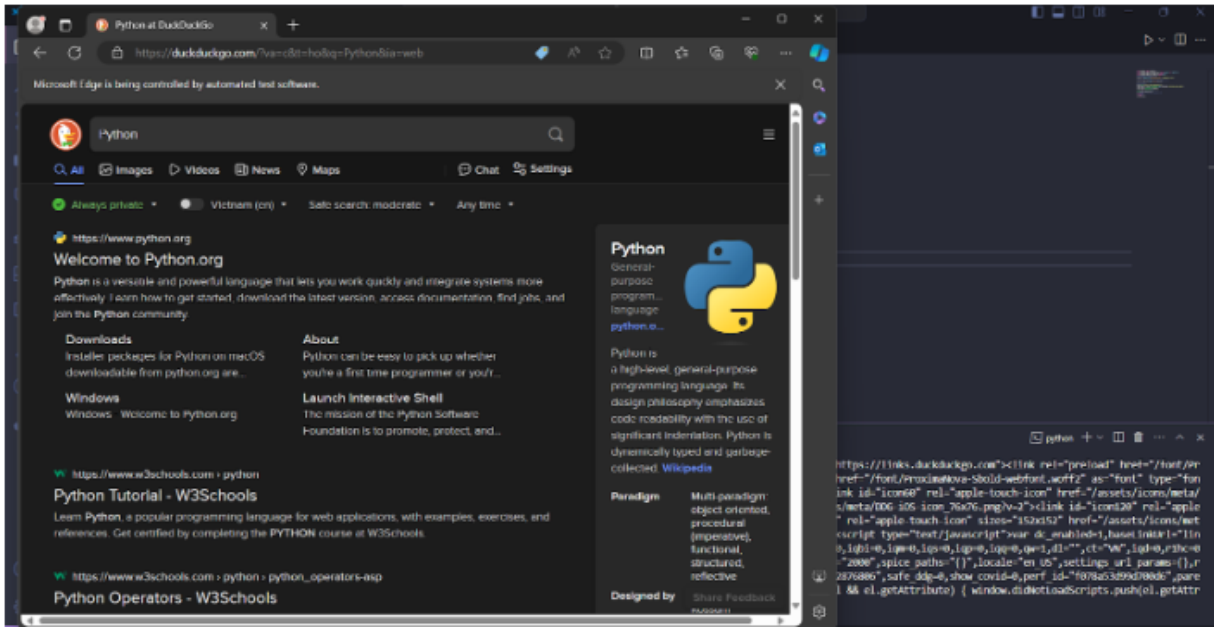
+) Tự động hóa Trình Duyệt: Selenium được sử dụng để tự động hóa các thao tác trên trình duyệt web như nhấn nút, điều hướng, hoặc điền vào các ô input.

+) Tối Ưu Hóa Crawl Động:

+) Chống chặn của Bot: Trong một số trường hợp, trang web có thể phát hiện và chặn các yêu cầu đến từ bot truy cập thông thường. Trong trường hợp này, Selenium có thể được sử dụng để mô phỏng hành vi người dùng thực sự trên trình duyệt, giúp tránh được các biện pháp chống chặn của bot.

(b) Sử dụng

Hình 1.27: Cấu hình selenium với scrapy



Hình 1.28: Giao diện hiển thị selenium

4. Kết luận

Tóm lại, cả Selenium và Splash đều là các công cụ mạnh mẽ trong lĩnh vực tự động hóa trình duyệt web, nhưng chúng có những ưu điểm và ứng dụng khác nhau.Sự lựa chọn giữa Selenium và Splash phụ thuộc vào

nhu cầu cụ thể của dự án. Đối với việc tự động hóa và kiểm thử các ứng dụng web phức tạp, Selenium thường là lựa chọn hàng đầu. Trong khi đó, nếu cần render JavaScript hoặc thu thập dữ liệu từ các trang web động một cách hiệu quả, Splash có thể là lựa chọn phù hợp hơn.

1.1.8 Tuần 10+11 (06/05 - 19/05)

Chủ đề bài tập lớn: Thu thập dữ liệu cá nhân trên nền tảng facebook

Kết quả của tuần 10:

1. Chốt đề tài: Thu thập dữ liệu cá nhân trên nền tảng facebook

KẾT LUẬN CHƯƠNG 1

Chương 1 tổng hợp lại toàn bộ quá trình học tập và nghiên cứu trong 11 tuần học của môn Thực tập cơ sở. Giai đoạn đầu tiên từ tuần 1 đến tuần 5, các chủ đề lý thuyết được lựa chọn một cách có khoa học nhằm phục vụ cho đề tài cuối kỳ của môn, đặc biệt là hai chủ đề tuần 4 và 5 là quy trình phát triển phần mềm và công cụ quản lý mã nguồn tạo cơ sở quan trọng cho giai đoạn sau. Giai đoạn sau từ tuần 6 + 7: đưa ra ý tưởng đề tài bài tập nhóm. Giao đoạn từ tuần 8 đến tuần 11, tìm hiểu về các các thức thu thập dữ liệu. Báo cáo hoàn thiện của sản phẩm đề tài cuối kỳ sẽ được trình bày chi tiết và kỹ càng hơn ở Chương 2.

CHƯƠNG 2. Báo cáo đề tài: Thu thập dữ liệu cá nhân trên nền tảng facebook

2.1 Giới thiệu đề tài

Trong thời đại mạng xã hội, việc quản lý và hiểu biết về mối quan hệ trong mạng lưới của chúng ta trở nên ngày càng quan trọng. Facebook, với hàng tỷ người dùng trên toàn thế giới, là một trong những nền tảng mạng xã hội lớn nhất và phức tạp nhất hiện nay. Tuy nhiên, với sự phát triển của mạng xã hội, việc quản lý và tương tác với một số lượng lớn bạn bè và người theo dõi trên Facebook trở nên khó khăn và đôi khi là không hiệu quả.

Để giải quyết vấn đề này, một công cụ sử dụng kỹ thuật cào dữ liệu từ trang cá nhân trên Facebook đã được phát triển. Bằng cách kết hợp các công nghệ có thể kể đến như Scrapy và Selenium, công cụ có thể thu thập danh sách bạn bè, thông tin về số lượng lượt tương tác và bình luận trên các bài đăng.Dữ liệu sẽ được tiến hành phân tích để xác định danh sách bạn bè ít tương tác nhất và trả về kết quả.

2.2 Xác định bài toán

Khi xem xét về trang cá nhân trên nền tảng Facebook, có thể nhận thấy hai loại tương tác chính từ người dùng khác:

+) Lượt tương tác: Đây là các hoạt động như "thích"(like) mà người dùng thực hiện đối với bài đăng của một người dùng khác trên Facebook. Lượt tương tác này thường là một chỉ báo về mức độ quan tâm hoặc sự đồng cảm của người dùng đối với nội dung đó. Trong khuôn khổ của bài toán, các lượt tương tác chỉ cần tính theo số lượng, không quan tâm đến thể loại

+) Lượt bình luận: Đây là các phản hồi cụ thể và tương tác trực tiếp từ người dùng đối với nội dung được đăng trên Facebook. Lượt bình luận thường cung cấp thông tin chi tiết hơn về ý kiến, suy nghĩ và phản ứng của người dùng đối với nội dung đó.

Dựa trên nhận xét trên, bài toán xác định người dùng ít tương tác nhất trên Facebook đối với trang cá nhân có thể được định nghĩa như sau:

+)Bài toán: Xác định danh sách bạn bè ít tương tác nhất trên Facebook dựa trên số lượng lượt tương tác trên trang cá nhân.

+) Xác định mục tiêu dữ liệu:

+)Danh sách bạn bè

+)Id facebook của bạn bè

+)Tên hiển thị

+)Lượt tương tác

+)Số lượt tương tác

+)Số lượt bình luận

+)Id của người tương tác

+)Id của người bình luận

+)Phương pháp thu thập dữ liệu: Sử dụng Scrapy kết hợp với Selenium để tự động thu thập dữ liệu. Scrapy giúp quản lý quá trình crawl dữ liệu, trong khi Selenium cho phép tương tác với trang web để thu thập thông tin. Khi thu thập được dữ liệu, có thể phân tích để xác định những người bạn ít tương tác nhất.

2.3 Công cụ chính

- Scrapy

- Mô tả: Scrapy là một framework mã nguồn mở mạnh mẽ được viết bằng Python, được sử dụng để crawl và trích xuất dữ liệu từ các trang web. Scrapy cung cấp các công cụ và thư viện giúp quản lý quá trình crawl dữ liệu một cách hiệu quả và có tổ chức.

- Chức năng trong dự án:

+)Tổ chức và quản lý quá trình crawl dữ liệu.

+)Xử lý và lưu trữ dữ liệu thu thập được.

- Selenium

- Mô tả : Selenium là một công cụ tự động hóa trình duyệt web, cho phép điều khiển các trình duyệt web từ chương trình máy tính. Selenium hỗ trợ nhiều ngôn ngữ lập trình, bao gồm Python, và có thể tương tác với các trang web theo cách mà người dùng thực hiện.

- Chức năng trong dự án:

+)Tự động đăng nhập vào tài khoản Facebook của người dùng.

+)Điều hướng và tương tác với các phần tử động trên trang web Facebook.

+)Trích xuất dữ liệu từ các phần tử HTML sau khi trang web đã được tải và hiển thị đầy đủ.

+)Xử lý các tình huống yêu cầu tương tác phức tạp, chẳng hạn như cuộn trang

2.4 Quy trình thực hiện

- Cấu hình spider

```python
import scrapy
from selenium.webdriver.common.by import By
import time
import datetime
from scrapy_selenium import SeleniumRequest
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from crawl_facebook.items import CrawlFacebookItem, CrawlFacebookReactItem, CrawlFacebookCommentItem
from selenium.webdriver.chrome.options import Options as ChromeOptions
from selenium.webdriver import Chrome
from selenium.webdriver.chrome.service import Service as ChromeService
from shutil import which
```

Hình 2.1: Khai báo thư viện

Khai báo các thư viện cần thiết đối với spider.

```python
class FacebookSpider(scrapy.Spider):
    name = "facebook"
    allowed_domains = ["www.facebook.com"]
    start_urls = [
        "https://www.facebook.com/",
        "https://www.facebook.com/profile.php?id=61560307553960&sk=friends",
        "https://www.facebook.com/profile.php?id=61560307553960"
    ]
    scroll_distance = 1080      You, 2 weeks ago • first commit

    def scroll_down_by(self, driver, distance):
        driver.execute_script(f"window.scrollBy(0, {distance});")

    def is_at_bottom(self, driver):
        return driver.execute_script("return window.innerHeight + window.scrollY >= document.body.scrollHeight")
```

Hình 2.2: Khởi tạo tham số spider

Khởi tạo các tham số cho spider.

+) allowed_domains:Tên miền mà spider được phép truy cập, spider chỉ thu thập dữ liệu từ các url thuộc tên miền trong danh sách này.

+)start_urls : Khai báo danh sách các url

+)def scroll_down_by: hàm này có tác dụng cuộn màn hình xuống một khoảng distance

+)def scroll_down_by:hàm này có tác dụng kiểm tra xem màn hình đã được cuộn hết hay chưa

```python
def __init__(self, *args, **kwargs):
    super(FacebookSpider, self).__init__(*args, **kwargs)
    chrome_path = which("chromedriver")
    chrome_options = ChromeOptions()
    chrome_options.add_argument("--incognito")
    chrome_options.add_argument("--headless")
    chrome_options.add_argument("--disable-gpu")
    chrome_options.add_argument("--window-size=1920,1080")
    chrome_options.add_argument("--no-sandbox")
    chrome_options.add_argument("--disable-dev-shm-usage")
    chrome_options.add_argument("--disable-extensions")
    chrome_options.add_argument("--disable-logging")
    chrome_options.add_argument('--log-level=3')
    chrome_service = ChromeService(chrome_path)
    self.driver = Chrome(service=chrome_service, options=chrome_options)
    self.login()
```

Hình 2.3: Cấu hình selenium

Cấu hình các tham số cơ bản cho selenium

+) -incognito:Chạy trình duyệt Chrome ở chế độ ẩn danh (incognito mode).

+) -headless:Chạy trình duyệt Chrome ở chế độ không có giao diện đồ họa (headless mode). Điều này giúp thực hiện các tác vụ mà không cần mở cửa sổ trình duyệt, tiết kiệm tài nguyên hệ thống.

+) -disable-gpu:Vô hiệu hóa việc sử dụng GPU. Tham số này thường được sử dụng cùng với chế độ headless để tránh các vấn đề liên quan đến hiển thị đồ họa.

+)-no-sandbox:Vô hiệu hóa chế độ sandbox của Chrome. Chế độ sandbox thường được sử dụng để tăng cường bảo mật, nhưng có thể gây ra các vấn đề khi chạy trong môi trường không có giao diện đồ họa

hoặc container.

+)-window-size:Thiết lập kích thước cửa sổ trình duyệt

+) -disable-extensions:Vô hiệu hóa tất cả các tiện ích mở rộng của Chrome.

+) –disable-logging:Vô hiệu hóa việc ghi log từ trình duyệt. Điều này giúp giảm bớt các thông tin log không cần thiết trong đầu ra.

```
def login(self):
    self.driver.get(self.start_urls[0])
    time.sleep(3)
    self.driver.implicitly_wait(3)
    input_account = WebDriverWait(self.driver, 3).until(
        EC.presence_of_element_located((By.XPATH, "//input[@id='email']"))
    )
    input_account.send_keys('boxdat123@gmail.com')
    input_password = self.driver.find_element(By.XPATH, "//input[@id='pass']")
    input_password.send_keys('123456789aA@')
    btn_sign = self.driver.find_element(By.XPATH, "//button[@name='login']")
    btn_sign.click()
```

Hình 2.4: Hàm login

Hàm login có tác dụng đăng nhập tự động vào facebook bằng cách sau khi selenium truy cập vào trang chủ của facebook, thông tin về tài khoản và mật khẩu sẽ được tự động điền và đăng nhập

```
def start_requests(self):
    yield SeleniumRequest(
        url=self.start_urls[1],
        wait_time=3,
        callback=self.parse_friend,
        dont_filter=True
    )
```

Hình 2.5: Hàm tạo request

Tạo ra một selenium request, bắt đầu khởi tạo quá trình cào dữ liệu. Hàm parse_friend sẽ được gọi đến và tiếp tục thực hiện trong chương trình

Hình 2.6: Hàm trích xuất bạn bè

Sau khi được gọi đến, parse_friend sẽ thực hiện lần lượt các thao tác:

+) Selenium driver sẽ truy cập vào địa chỉ chứa danh sách bạn bè, cụ thể ở đây là start_url[1] để thực hiện quá trình cào dữ liệu

+)Bên trong hàm có một vòng lặp kiểm tra xem trang đã được cuộn hết chưa

+)Nếu chưa dữ liệu sẽ được xác định bằng phương thức Xpath.Dữ liệu sẽ được cào ra và định nghĩa bởi các items, sau đó được chuyển đến lớp pipeline để xử lý. Nếu xảy ra trường hợp không xác định được phần tử, hàm sẽ ném ra một ngoại lệ và bắt đầu cuộn trang một khoảng bằng với tham số scroll_distance được khai báo ở trên. Sau đó sẽ tiếp tục quay lại với vòng lặp và kiểm tra xem đã cuộn tới cuối trang chưa

+)Khi đã cào hết danh sách bạn bè, một selenium request sẽ được tạo ra để điều hướng tiếp tục quá trình cào dữ liệu, hàm parse_personal_page sẽ được gọi đến, thêm vào đó là địa chỉ url trang web tiếp theo được điều hướng tới

Hình 2.7: Hàm trích xuất tương tác 1

Hình 2.8: Hàm trích xuất tương tác 2

Sau khi được gọi đến từ selenium request, hàm parse_personal_page, sẽ thực hiện quá trình trích xuất dữ liệu từ trang cá nhân. Bên trong hàm tồn tại một vòng lặp , kiểm tra xem đã cuộn đến cuối trang chưa

+) Nếu chưa, dữ liệu sẽ được xác định bằng phương thức Xpath.Dữ liệu sẽ được cào ra và định nghĩa bởi các items, sau đó được chuyển đến lớp pipeline để xử lý. Nếu xảy ra trường hợp không xác định được phần tử, hàm sẽ ném ra một ngoại lệ và bắt đầu cuộn trang một khoảng bằng với tham số heigh được định nghĩa trong hàm. Sau đó sẽ tiếp tục quay lại với vòng lặp và kiểm tra xem đã cuộn tới cuối trang chưa.

+) Đến khi trang đã được cuộn hết, dữ liệu đã được cào hoàn tất, chương trình sẽ kết thúc

- Lớp items

```python
# https://docs.scrapy.org/en/latest/topics/items.html

import scrapy
class CrawlFacebookItem(scrapy.Item):
    # define the fields for your item here like:
    name = scrapy.Field()
    idUser = scrapy.Field()
    time = scrapy.Field()
    react = scrapy.Field()
    comment = scrapy.Field()
class CrawlFacebookReactItem(scrapy.Item):
    idUser = scrapy.Field()
class CrawlFacebookCommentItem(scrapy.Item):
    idUser = scrapy.Field()
```

Hình 2.9: Lớp Item

Định nghĩa các items sẽ được sử dụng trong suốt quá trình cào dữ liệu

- Lớp pipeline

```python
import pymongo
import datetime
import re
from scrapy.exceptions import DropItem
from selenium import webdriver
from selenium.webdriver.edge.options import Options as EdgeOptions
from selenium.webdriver import Edge
from selenium.webdriver.chrome.service import Service as ChromeService
from selenium.webdriver.edge.service import Service as EdgeService
from shutil import which
import time
from crawl_facebook.items import CrawlFacebookItem, CrawlFacebookReactItem, CrawlFacebookCommentItem
```

Hình 2.10: Khai báo thư viện

Định nghĩa và khai báo các thư viện cần thiết đối với lớp pipeline

```python
# See https://docs.scrapy.org/en/latest/topics/item-pipeline.html
ITEM_PIPELINES = {
   "crawl_facebook.pipelines.CrawlFacebookPipeline": 300,
}



# Set settings whose default value is deprecated to a future-proof value
REQUEST_FINGERPRINTER_IMPLEMENTATION = "2.7"
TWISTED_REACTOR = "twisted.internet.asyncioreactor.AsyncioSelectorReactor"
FEED_EXPORT_ENCODING = "utf-8"
MONGO_URI = 'mongodb://localhost:27017'
MONGO_DATABASE = 'friend_facebook'
MONGO_COLLECTION = 'list_friend_react'
LOG_LEVEL = 'ERROR'

```

Hình 2.11: Khai báo biến môi trường

Khai báo các biến môi trường đối với quá trình cào dữ liệu

```python
def __init__(self, mongo_uri, mongo_db, mongo_collection):
    self.mongo_uri = mongo_uri
    self.mongo_db = mongo_db
    self.mongo_collection = mongo_collection
@classmethod
def from_crawler(cls, crawler):
    return cls(
        mongo_uri=crawler.settings.get('MONGO_URI'),
        mongo_db=crawler.settings.get('MONGO_DATABASE'),
        mongo_collection=crawler.settings.get('MONGO_COLLECTION')

    )
def open_spider(self, spider):      You, 2 weeks ago • first commit
    self.client = pymongo.MongoClient(self.mongo_uri)
    self.db = self.client[self.mongo_db]
```

Hình 2.12: Khởi tạo các kết nối đến cơ sở dữ liệu

Khởi tạo các kết nối với cơ sở dữ liệu

```python
def process_item(self, item, spider):
    edge_path = which("msedgedriver")
    edge_options = EdgeOptions()
    edge_options.add_argument("--incognito")
    edge_options.add_argument("--headless")
    edge_options.add_argument("--disable-gpu")
    edge_options.add_argument("--log-level=3")
    edge_options.add_argument("--no-sandbox")
    edge_options.add_argument("--disable-dev-shm-usage")
    edge_options.add_argument("--disable-extensions")
    edge_options.add_argument("--disable-logging")
    edge_service = EdgeService(edge_path)
    driver_edge = Edge(service=edge_service, options=edge_options)
    driver_edge.get(item.get('idUser'))
    time.sleep(3)
    driver_edge.implicitly_wait(3)
    html = driver_edge.page_source
    match = re.search(r'"userID":"(\d+)"', html)
    if match:
        user_id = match.group(1)
    else:
        user_id = None
    if user_id:
        item['idUser'] = user_id
    driver_edge.quit()
```

Hình 2.13: Xử lý dữ liệu

Sử dụng selenium trích xuất user id. Sử dụng selenium driver để truy cập vào trang cá nhân để lấy ra user id của người dùng.

```python
    if 'time' in item:
        try:
            item['time'] = datetime.datetime.strptime(item['time'], '%Y-%m-%d %H:%M:%S')
        except ValueError as e:
            raise DropItem(f"Failed to parse 'time' field: {e}")
```

Hình 2.14: Định dạng kiểu dữ liệu

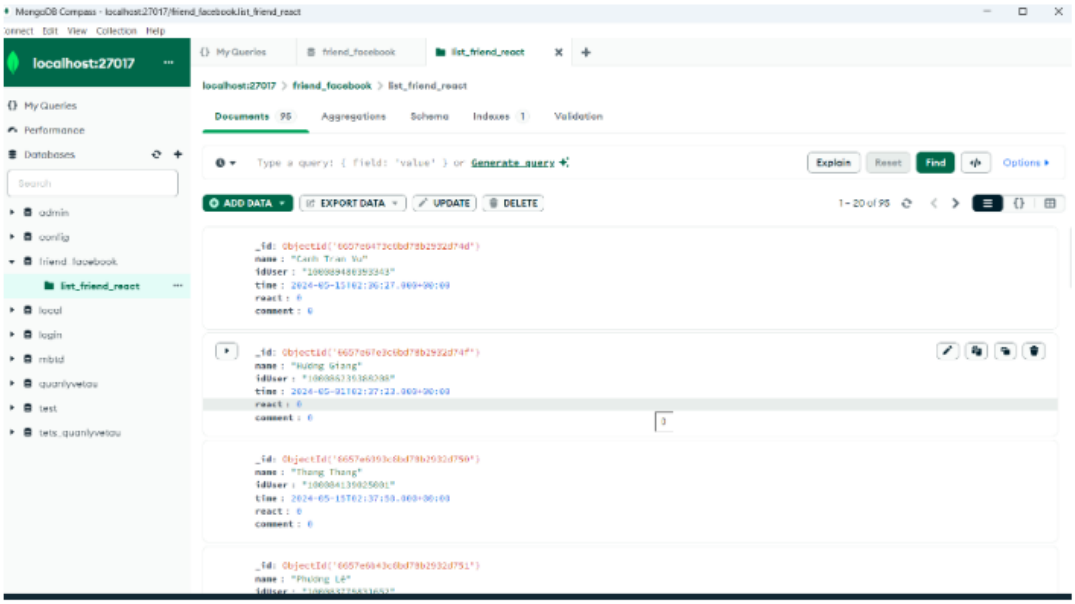
Định dạng kiểu dữ liệu thời gian sau khi items nhận được items từ spider

```python
def process_item(self, item, spider):
    if 'time' in item:
        except ValueError as e:
    if isinstance(item, CrawlFacebookItem):
        try:
            existing_item = self.db[self.mongo_collection].find_one({'idUser': item.get('idUser')})
            if existing_item:
                updates = {}
                if existing_item.get('name') != item.get('name'):
                    updates['name'] = item.get('name')
                updates['react'] = 0
                updates['comment'] = 0
                if updates:
                    self.db[self.mongo_collection].update_one(
                        {'_id': existing_item['_id']},
                        {'$set': updates}
                    )
            else:
                self.db[self.mongo_collection].insert_one(dict(item))
                return item
        except:
            raise DropItem(f"Failed to update item in MongoDB: {e}")
    elif isinstance(item, CrawlFacebookReactItem):
      #drop database
        existing_item = self.db[self.mongo_collection].find_one({'idUser': item.get('idUser')})
        if existing_item:
            self.db[self.mongo_collection].update_one(
                {'_id': existing_item['_id']},
                {'$inc': {'react': 1}}
            )
        else:
            raise DropItem(f"Missing one of the fields in {item}")
    elif isinstance(item, CrawlFacebookCommentItem):
        existing_item = self.db[self.mongo_collection].find_one({'idUser': item.get('idUser')})
        if existing_item:
            self.db[self.mongo_collection].update_one(
                {'_id': existing_item['_id']},
                {'$inc': {'comment': 1}}
            )
        else:
            raise DropItem(f"Missing one of the fields in {item}")
    else:
        raise DropItem(f"Unknown item type: {type(item)}")
```

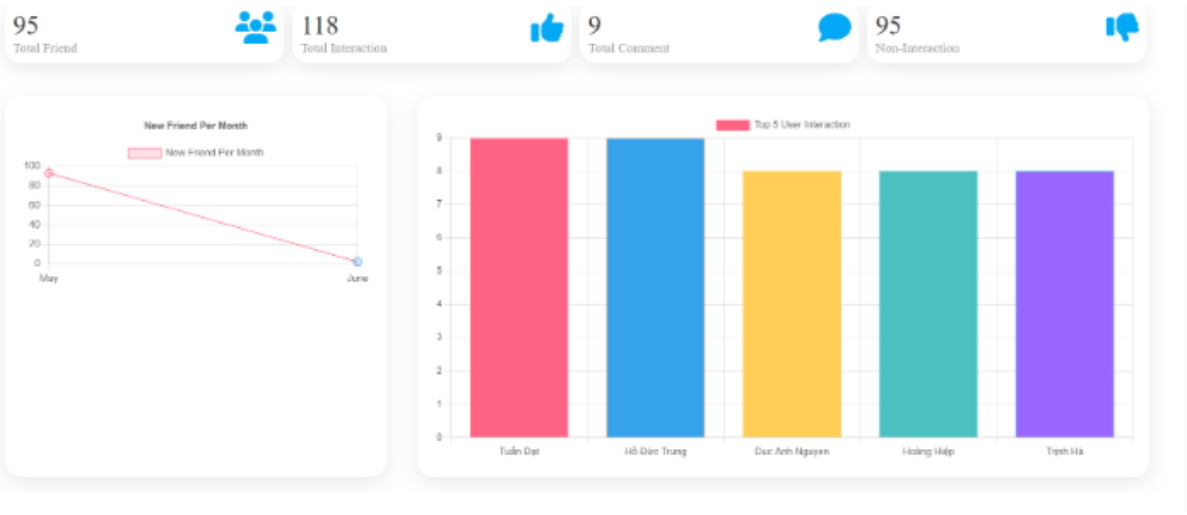
Hình 2.15: Lưu trữ dữ liệu

Kiểm tra các trường hợp của dữ liệu để đưa vào cơ sở dữ liệu.

- Kết quả



Hình 2.16: Dữ liệu sau khi cào được lưu vào cơ sở dữ liệu

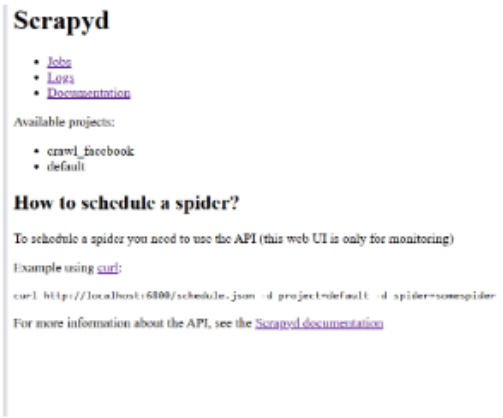


Hình 2.17: Giao diện hiển thị dữ liệu

STT	ID	NAME	INTERACTION	COMMENT
1	100089480393343	Canh Trao Vu	0	0
2	100086239388208	Huong Giang	0	0
3	100084139025001	Thang Thang	0	0
4	100083779831652	Phương Lê	0	0
5	100072411849964	Huy Init	0	0
6	100065669885010	Thùy Linh	0	0
7	100063452289304	Hữu Soo	0	0
8	100062908186571	Nguyệt Ánh	0	0

Hình 2.18: Danh sách user id ít tương tác nhất

- Lên lịch



Hình 2.19: Scapyd

Đối với việc lên lịch chạy tự động, công cụ sẽ được triển khai trên hai nền tảng là Scrapyd và Scrapydweb.

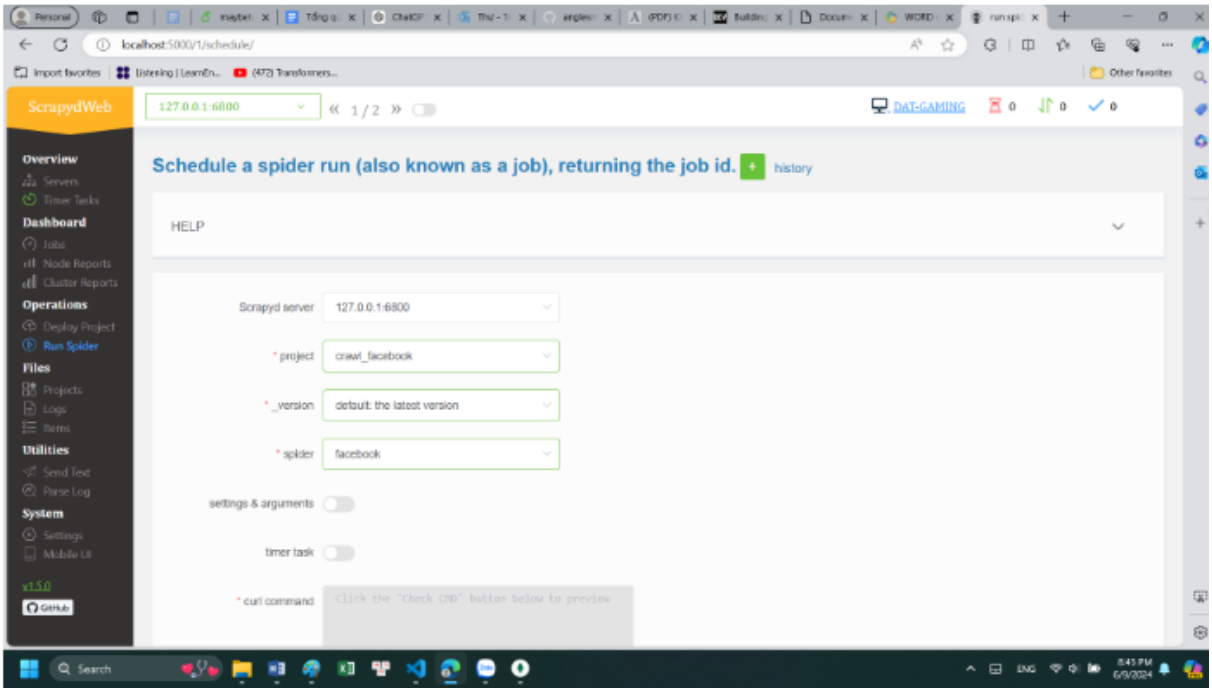
+)Scrapyd cung cấp API để spider có thể được triển khai từ xa. Điều này cho phép người dùng viết các tác vụ lên lịch tự động trên hệ thống của mình để gọi các API này theo định kỳ.

+)Một giao diện web được cung cấp bởi Scrapydweb để việc quản lý và giám sát các dự án và công việc Scrapyd trở nên dễ dàng. Qua giao diện này, lịch chạy tự động cho spider có thể được tạo và quản lý một cách thuận tiện.

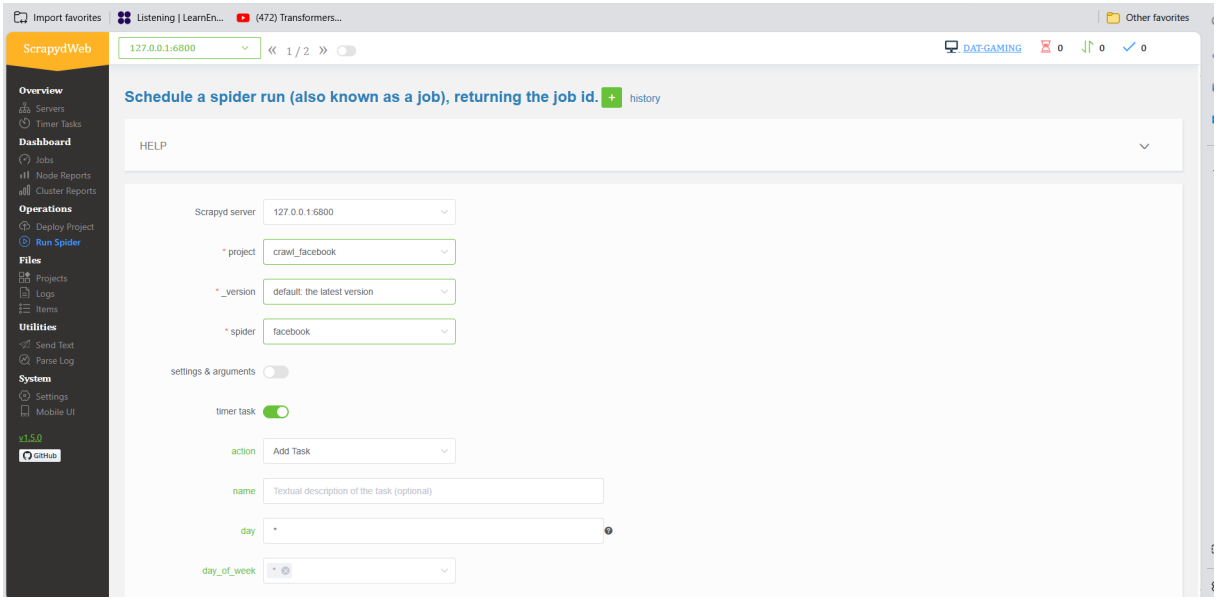
```
# https://scrapyd.readthedocs.io/en/latest/deploy.html

[settings]
default = crawl_facebook.settings


# nhớ phải đặt tên project để deploy
[deploy:local]
url - http://localhost:6800/
project=crawl_facebook
DOWNLOAD_DELAY = 600
```

Hình 2.20: Cấu hình spider trong scrapy.cfg

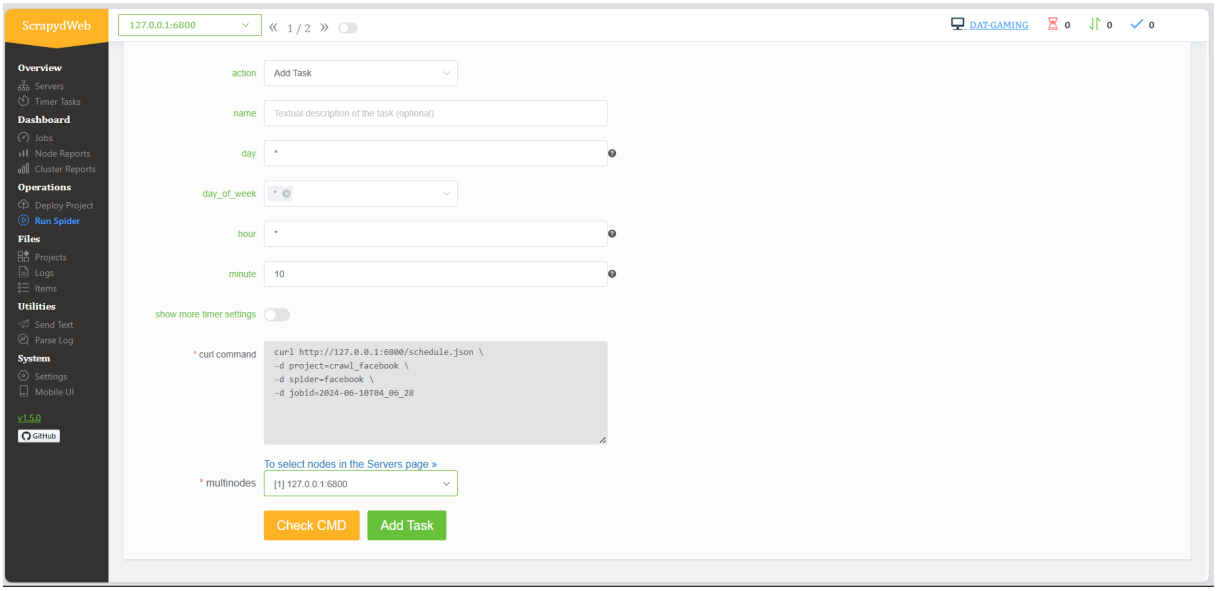
Cấu hình spider để triển khai đối với scrapyd



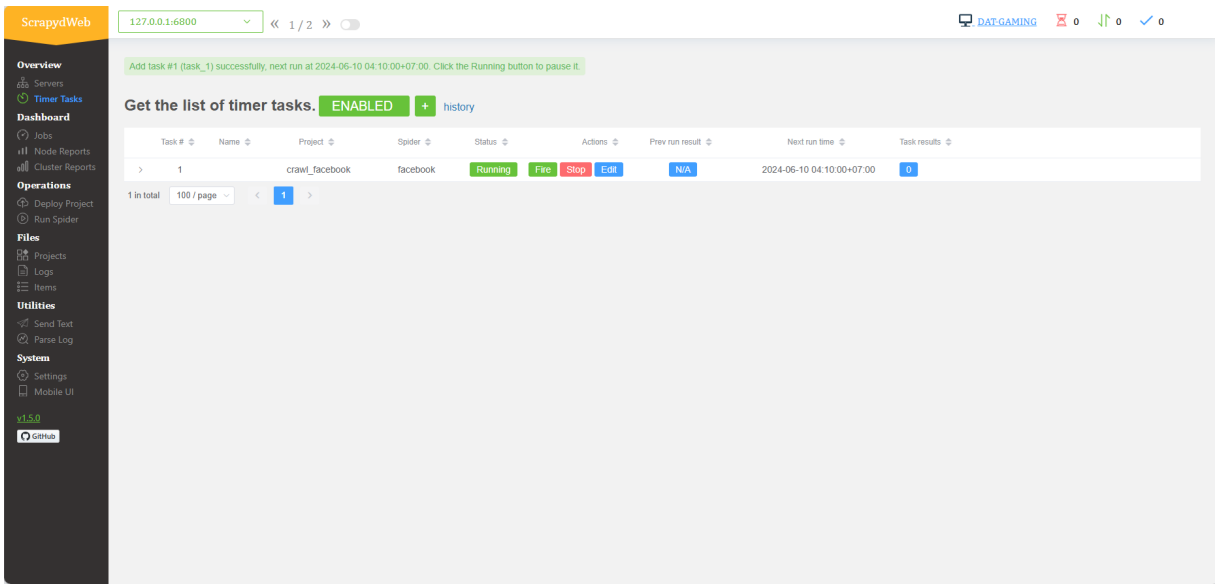
Hình 2.21: Giao diện scrapydweb



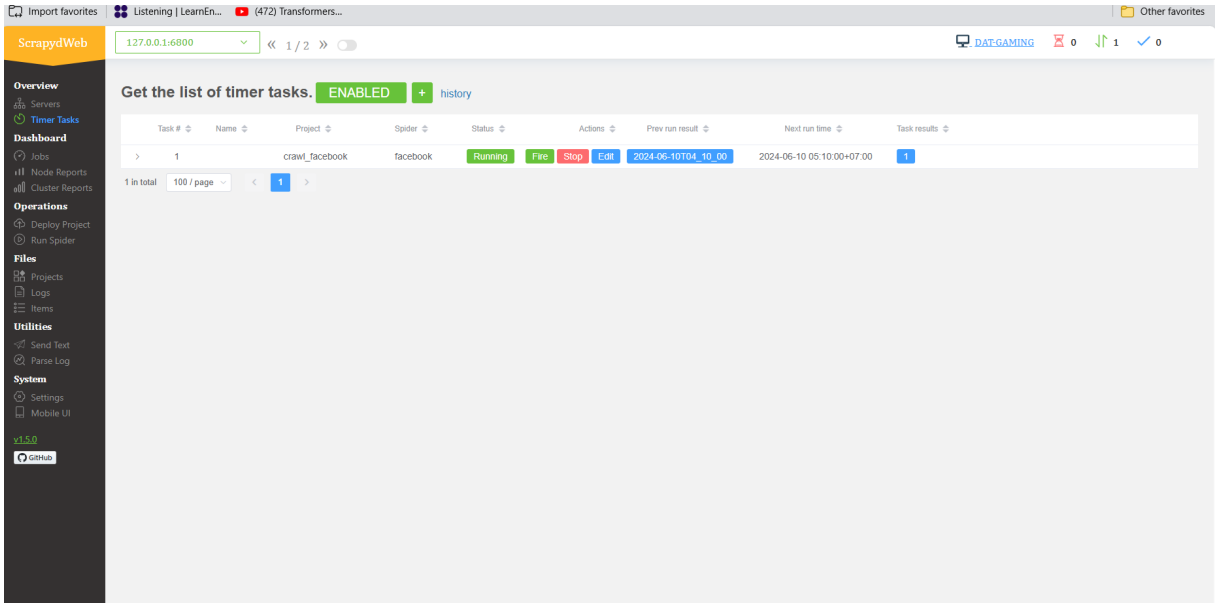
Hình 2.22: Lựa chọn spider



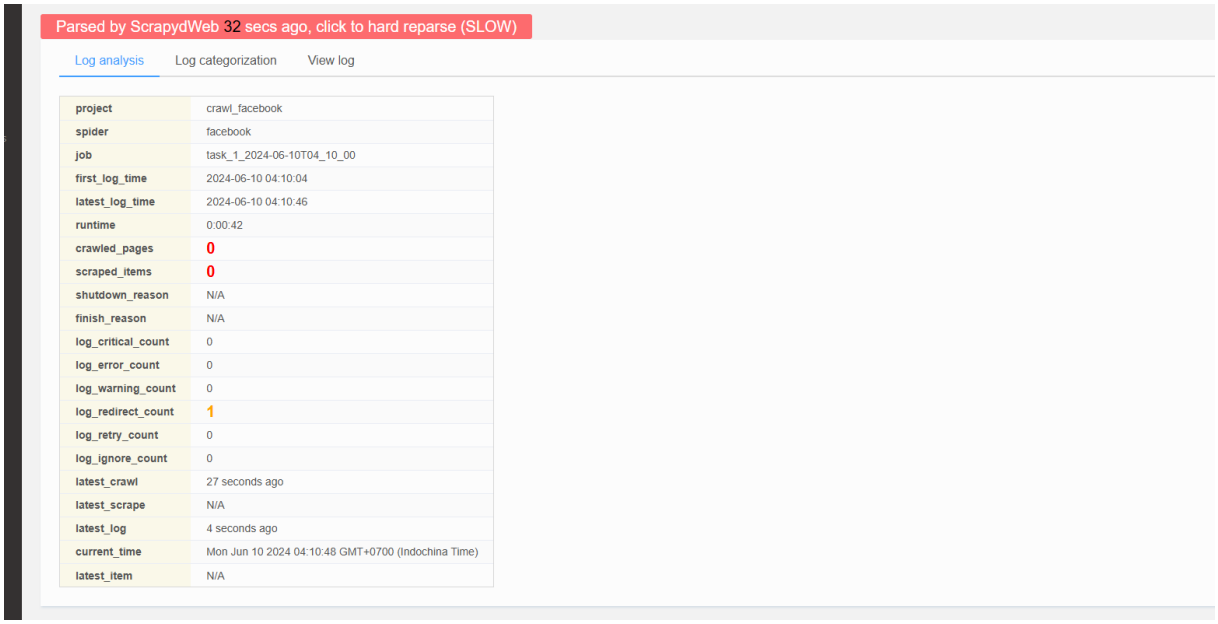
Hình 2.23: Cấu hình thời gian chạy



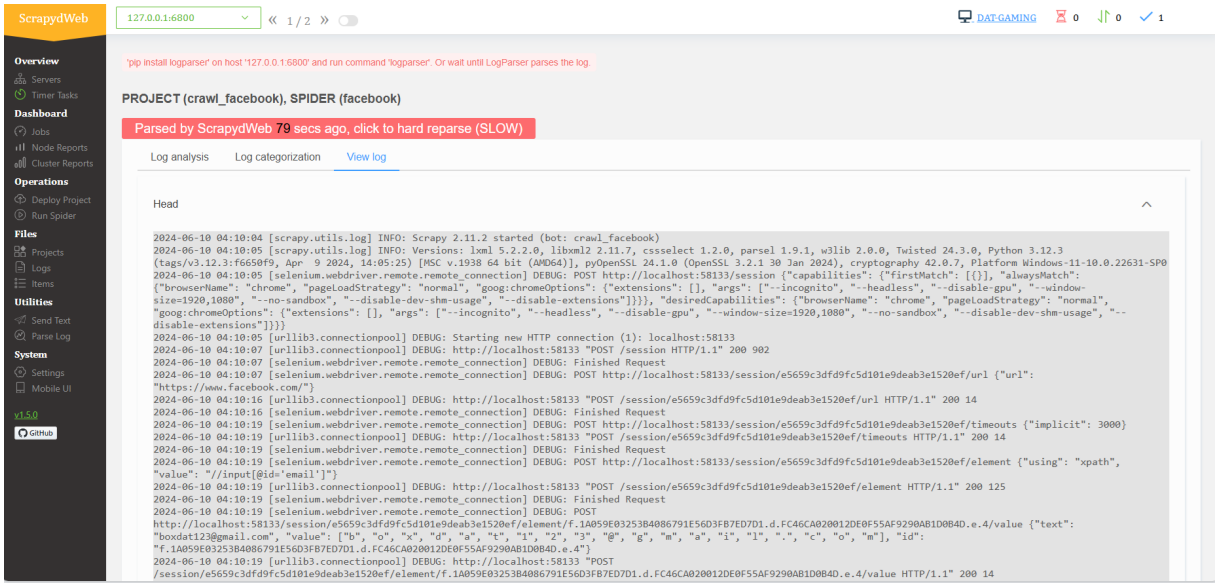
Hình 2.24: Lên lịch chạy thành công



Hình 2.25: Spider khởi chạy

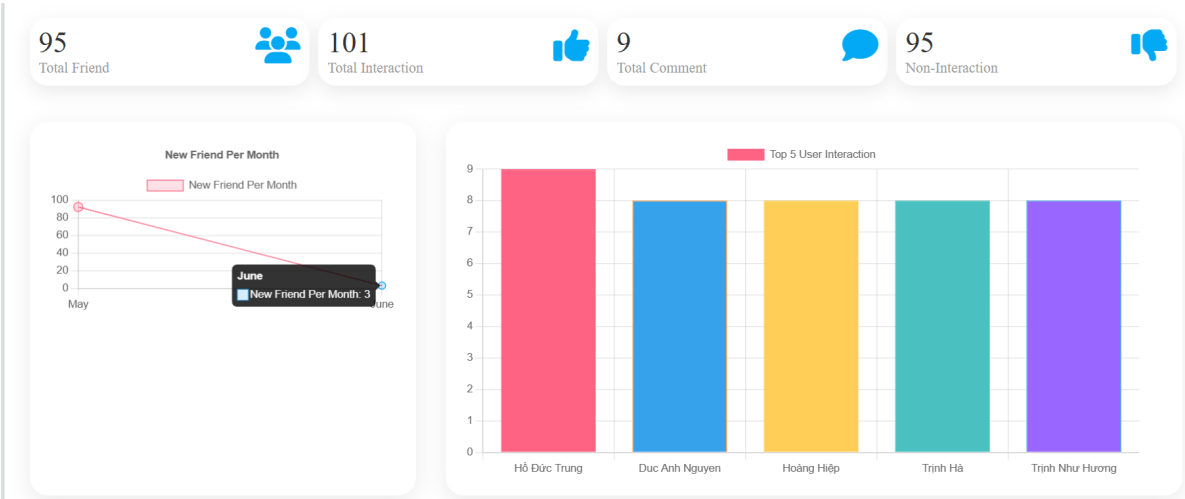


Hình 2.26: Thông báo hiển thị



Hình 2.27: Log spider

Spider đã được lên lịch và chạy thành công. Dữ liệu thay đổi sẽ được update trong cơ sở dữ liệu.



Hình 2.28: Dữ liệu hiển thị

Chúng ta có thể thấy dữ có một chút sự thay đổi đến từ danh sách bạn bè và lượt tương tác

CHƯƠNG 3. KẾT LUẬN

3.1. Đánh giá:

3.1.1. Ưu điểm:

- Cá nhân đã hoàn thành đề tài đưa ra ở tuần 10 - 11
- Công cụ có thê đưa ra danh sách những bạn bè ít tương tác, nhờ vào đó có thể giúp cá nhân người dùng tăng trải nghiệm sử dụng nền tảng mạng xã hội facebook

3.1.2. Hạn chế:

- Do sử dụng đồng thời 2 driver của selenium nên xảy ra tình trạng chậm, chưa thể cải thiện được hiệu suất tốt nhất

3.2. Hướng khắc phục và phát triển trong tương lai:

- Sự hạn chế của đề tài nằm ở việc sử dụng cùng lúc 2 driver của Selenium, dẫn đến tình trạng chậm trễ và không thể đạt hiệu suất tốt nhất. Để khắc phục vấn đề này và phát triển trong tương lai, đội ngũ nghiên cứu sẽ tìm hiểu về các chính sách bảo mật và mở rộng các phương pháp thu thập dữ liệu khác như sử dụng API.

TÀI LIỆU THAM KHẢO

[1] Scrapy Contributors, *Scrapy documents*. [Online]. Available: `https://docs.scrapy.org/en/latest/` (visited on 10/06/2024).

[2] Scrapy Contributors , *Scrapyd Document*. [Online]. Available: `https://scrapyd.readthedocs.io/en/latest/` (visited on 10/06/2024).

[2] Christian Clauss LXL , *Scrapydweb Document*. [Online]. Available: `https://github.com/my8100/scrapydweb` (visited on 10/06/2024).