**ĐẠI HỌC HUẾ**

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

ĐỒ ÁN MÔN HỌC

MÔN NGÔN NGỮ LẬP TRÌNH PYTHON



Sinh viên: Trần Lý Bửu

Huế, ngày 4 tháng 2 năm 2021

***Mục lục***

[**Chương 1: Cơ Sở Lý Thuyết** 1](#_Toc63018069)

[**1.1 - Giới thiệu về hệ thống Web Crawler** 2](#_Toc63018070)

[**1.2 – Cấu trúc của một trang web - HTML** 5](#_Toc63018071)

[**1.3 – Python và các thư viện hỗ trợ** 6](#_Toc63018072)

[**Chương 2: Thiết Kế Hệ Thống Web Crawler** 9](#_Toc63018073)

[**2.1 – Kiến trúc hệ thống** 10](#_Toc63018074)

[**2.2 – Lưu đồ thuật toán** 11](#_Toc63018075)

[**2.3 – Kết quả thu được** 11](#_Toc63018076)

[**Chương 3: Kết Luận Về Đồ Án và Tài Liệu Kham Khảo** 13](#_Toc63018077)

**Github:** <https://github.com/tranlybuu/Project_HUET_Web_Crawler>

# **Chương 1 Cơ Sở Lý Thuyết**

Trả lời cho những câu hỏi:

* Web Crawler / Spider là gì ?
* Nguồn gốc Web Crawler ở đâu ?
* Một trang web được hình thành từ đâu ?
* HTML bắt nguồn từ đâu ?
* Ngôn ngữ lập trình Python giải quyết được những vấn đề nào ?
* Các thư viện nào có thể hỗ trợ ?

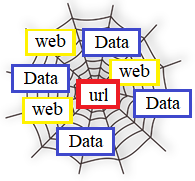
## **1.1 - Giới thiệu về hệ thống Web Crawler**

Google, Bing, Yahoo,… là những công cụ tìm kiếm rất nổi tiếng và được sử dụng phố biến hiện nay từ người dùng máy tính, laptop, điện thoại, máy tính bảng,…. Để những công cụ tìm kiếm này hoạt động thì dữ liệu là một điều không thể thiếu. Vậy những nhà phát triển đã lấy dữ liệu đó từ đâu ? Làm thế nào để họ có thể lấy những dữ liệu cho người dùng tìm kiếm ?

Những công cụ tìm kiếm trên đang hoạt động dựa theo quy tắc là tải những đường dẫn website về, lọc ra thông tin cần thiết, lưu trữ lại, tiếp theo là dựa vào thông tin cần tìm kiếm của khách hàng để lọc ra những trang web có chứa các từ khóa, chủ đề, thông tin liên quan rồi hiển thị ra cho khách hàng lựa chọn. Quá trình tải về và lọc thông tin để lưu trữ hay truy xuất chính là Web Crawler.

Web Crawler được lên ý tưởng và phát triển đầu tiền bởi Brian Pinkerton. Đây là một phần mềm được thiết kế nhằm mục đích có thể duyệt hầu hết website trên World Wide Web một cách có hệ thống để thu thập thông tin cần thiết về trang web đó và cung cấp cho công cụ tìm kiếm. Điều này mang lại khả năng lưu chỉ mục những trang web đó vào bộ cơ sở dữ liệu của công cụ tìm kiếm. Đồng thời, giúp những công cụ tìm kiếm đó có thể tìm ra các đánh giá chính xác nhất về một website đã được thu thập dữ liệu.

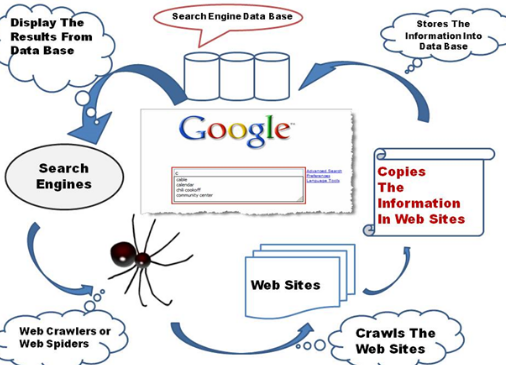
Web Crawler còn có những tên gọi khác như là spider, worm, ant,  robot, bot,…. Nhưng tên gọi thông thường nhất vẫn là crawler và mỗi tên gọi đều có những lý do của chúng.

Spider là một cách gọi hình tượng hóa, cái tên này dựa theo nguyên lý hoạt động và lưu trữ thông tin rất giống với hoạt động của con nhện. Bắt đầu ở một website bất kỳ và Spider sẽ len lỏi vào hết các ngóc ngách ở trong trang đó để lần lượt truy cập vào những liên kết có trên trang. Tiếp theo nó sẽ đánh dấu những liên kết đã từng truy cập trước đó để nối các trang có liên kết với trang gốc giống như con nhện đang tạo những sợi tơ liên kết các trang lại với nhau. Chỉ đơn giản bắt đầu từ một website, Spider có thể liên kết thêm rất nhiều website khác để hình thành một mạng lưới chằng chịt giống như một mạng nhện đích thực.

Ant là một cách gọi tương tự spider, cũng dựa trên cách thức lưu trữ thông tin của web và các hoạt động của Web Crawler. Sau mỗi lần di chuyển, con kiến sẽ tiết ra dung dịch pheromone với mục đích lưu lại vết đường mà nó đã từng đi qua. Với việc đánh dấu liên kết này trên web cũng giống với việc tạo liên kết tơ nhện của Spider.

Crawler là cách gọi dựa theo chức năng của nó, tên gọi này đã mô tả những hành động truy cập, thu thập dữ liệu trên những website giống như một người hoặc con bọ đang bò trườn qua những trang đó.

Bot (hay Internet Bot) là một loại phần mềm vận hành một cách tự động hóa trên Internet và robot có thể thực hiện những công việc đơn giản, lặp đi lặp lại theo một cách có hệ thống cho người sử dụng. Tóm tắt lại, Web Crawler là một tập hợp con của Internet Bot.

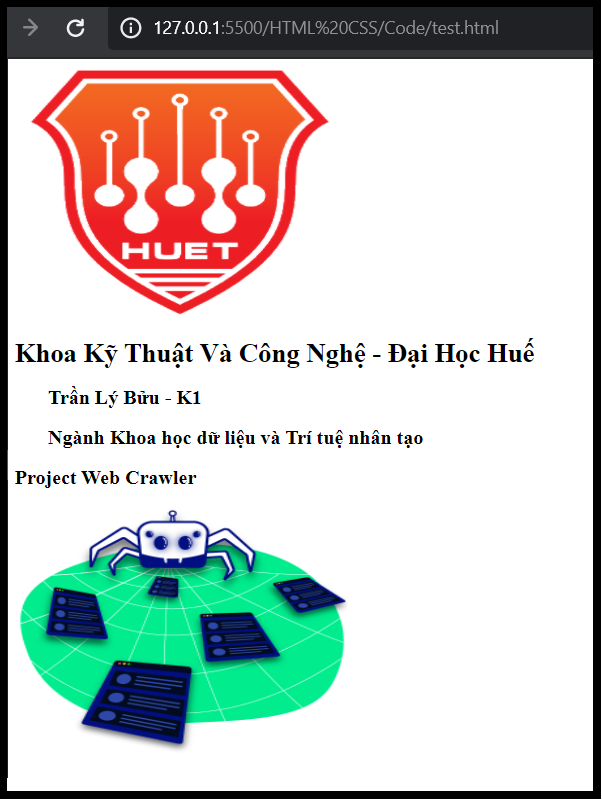


Sau khi đã hiểu khái quát về Web Crawler thì bây giờ chúng ta sẽ tìm hiểu thêm về những gì mà Web Crawler có thể mang lại cho người sử dụng. Như ban đầu tôi có đề cập, hầu hết các công cụ tìm kiếm phổ biến hiện nay vẫn đang sử dụng mô hình Web Crawler như DuckDuckGo sử dụng DuckDuckBot, Yandex sử dụng Yandex Bot, Yahoo! sử dụng Yahoo! Slurp,…. Những công cụ tìm kiếm lớn hơn như Bing, Google thì việc Web Crawler được chia nhỏ ra thành nhiều bot với chức năng riêng biệt. Điều này sẽ giúp việc thu thập và truy xuất dữ liệu diễn ra nhanh hơn và chuyên nghiệp hơn. Bing có một Internet Bot chuyên thu thâp thông tin web là Bing Bot, thu thập hình ảnh hay video là MSNBot-Media và thu thập các bài báo, những đánh giá thì có Bing Preview. Google có Googlebot dùng để thu thập thông tin từ cách thiết bị di động và máy tính, Googlebot Images thu thập hình ảnh, Googlebot Videos thu thập video, Googlebot News thu thập các bài báo và AdsBot được dùng để thu thập và chạy quảng cáo. Vậy tại sao Web Crawler lại quan trọng đối với SEO (Search Engine Optimization nghĩa là tối ưu hóa cho công cụ tìm kiếm)? Web Crawler yêu cầu các trang web phải cho phép truy cập và thu thập dữ liệu bởi vì thu thập dữ liệu là cách đầu tiên mà các công cụ tìm kiếm có thể lọc ra các từ khóa về trang web. Từ đó, bạn có thể xem xét và chọn lọc các từ khóa để trang web bạn xuất hiện nhiều trong mỗi lần người khác sử dụng công cụ tìm kiếm.

Tuy nhiên, hầu hết những công cụ tìm kiếm hiện nay đều không cho phép bạn thu thập thông tin các trang web không giới hạn. Điều này sẽ giúp những trang web không bị quá tải khi có quá nhiều Web Crawler đang hoạt động. Nếu bạn không muốn trang web của bạn bị quá nhiều Internet Bot thu thập dữ liệu dẫn đến hiện tượng bị nghẽn, quá tải, bạn có thể cài đặt và giới hạn việc thu thập thông tin từ Bot. Nhu cầu thu thập thông tin được đánh giá là mức độ quan tâm về trang web của Google và người dùng. Biết được cách thức hoạt động, bạn có thể lợi dụng nó để chọn lọc từ khóa giúp các Internet Bot có thể tăng mức độ tiếp cận của người dùng đến với trang web của bạn hoặc chặn những trình thu thập dữ liệu web có những thông tin nhạy cảm hoặc không liên quan với những từ khóa của người dùng tìm kiếm.

## **File:CLISK 1334741064.jpg - Wikimedia Commons1.2 – Cấu trúc của một trang web - HTML**

Để biết được cách hoạt động của Web Crawler, chúng ta cần biết về cấu trúc của một trang web. Website là một hoặc tập hợp nhiều trang web nằm trong tên miền trên World Wide Web của Internet. Một trang web được cấu tạo từ tập tin HTML hay XHTML và có thể truy nhập thông qua giao thức HTTP. Được phát triểu bởi WHATWG và phát hành năm 1993, HTML (Hypertext Markup Language hay Ngôn ngữ Đánh dấu Siêu văn bản) là ngôn ngữ đánh dấu đã được thiết kế nhằm tạo ra trang web trên World Wide Web. Kết hợp với CSS và JavaScript, HTML được xem như một ngôn ngữ quan trọng về lĩnh vực website. Sau khi trở thành chuẩn mực của Internet do [World Wide Web Consortium](https://vi.wikipedia.org/wiki/W3C) (W3C) duy trì, HTML được các nhà phát triển thay thế bằng XHTML và phiên bản mới nhất hiện nay là HTML5. Nếu chúng ta nói HTML là ngôn ngữ lập trình thì sai nhưng chúng ta có thể đọc được nội dung toàn bộ trang web thông qua HTML. HTML là một ngôn ngữ đánh dấu bằng thẻ (Markup Language), nói một cách dễ hiểu là bạn sử dụng HTML để đánh dấu các tài liệu bằng các thẻ (tag). Trang web có thể cấu trúc lại và hiển thị nội dung. Tuy nhiên, HTML dùng để đánh dấu trình bày bây giờ không còn được tối ưu mà nó được thay thế bằng CSS hay JavaScript.

Cấu trúc của HTML bao gồm rất nhiều thẻ như <html>, <title>, <head>, <body>,…. Vì những thứ bạn thu thập dữ liệu đều có trong file code HTML và không có trong CSS hay JavaScript nên nếu bạn chưa có kiến thức cơ bản về HTML nhưng muốn thực hiện Web Crawler thì bạn cần biết thẻ <title> chứa chủ đề trang web, <h1> đến <h6> chứa các tiêu đề, <p> chứa nội dung văn bản, <img> chứa đường dẫn hình ảnh, <a> chứa các đường dẫn Chỉ với vài dòng lệnh HTML đơn giản là bạn có thể tạo ra một trang offline mà chưa cần qua chỉnh sửa hay cấu trúc lại bằng CSS và JavaScript. Bạn xem một trang bên này sẽ được hình thành từ bao nhiêu dòng lệnh ?

Chưa tới 20 dòng lệnh và một IDE (Integrated Development Environment - phần mềm cung cấp cho lập trình viên phát triển) là bạn có thể viết một file HTML để dễ dàng tìm hiểu thêm về HTML.

Link code: <https://github.com/tranlybuu/Project_HUET_Web_Crawler/blob/master/demo_html_code.html>

## **1.3 – Python và các thư viện hỗ trợ**



Được tạo ra bởi [Guido van Rossum](https://vi.wikipedia.org/w/index.php?title=Guido_van_Rossum&action=edit&redlink=1) vào năm 1991 và được phát triển bởi nhóm Python Software Foundation, Python đã được ra đời và nổi tiếng với nhiều biệt danh khác nhau như “ngôn ngữ lập trình bậc cao” hay “ngôn ngữ lập trình hướng đối tượng”,…. Đây còn là một ngôn ngữ phiên dịch nên cú pháp ngắn gọn, dễ học, dễ nhớ và dễ đọc nhưng đổi lại là tốc độ không thể bằng C++,.... Python được phát triển ban đầu trên nền Unix và được mở rộng sang những hệ điều hành phổ biến hiện nay như MacOS, Linux, Window và các hệ điều hành khác những thuộc họ Unix. Bắt đầu với phiên bản 1.x từ năm 1990, Python được phát triển tiếp với phiên bản 2.x vào năm 2000. Phiên bản mới nhất của Python chính là 3.x (còn gọi là Python 3000 hoặc Py3K). Dòng 3.x này hoàn toàn khác với 2.x nhưng hiện tại đã có rất nhiều công cụ giúp bạn có thể chuyển đổi từ 2.x sang 3.x một cách dễ dàng. Tại sao lại có sự thay đổi lớn khi chuyển từ 2.x sáng 3.x vậy? Sự thay đổi lớn nhất nằm ở nguyên tác chủ đạo phát triển. Thay đổi cách làm cũ ở 2.x và làm lại 3.x nhằm tránh trùng lặp chức năng của các cú pháp Python.

Python là ngôn ngữ có đặc điểm tương tự thể thơ Haiku. Mọi cú pháp đều dễ học, dễ hiểu, ngắn gọn, mã nguồn mở và bố cục trực quan. Python có phân biệt chữ hoa và chữ thường, sử dụng các từ khóa ở dạng chữ giống như C/C++ điều này giúp hạn chế các ký hiệu và người đọc có thể dễ hiểu mà không cần tìm hiểu quá chuyên sâu nhưng vẫn có thể sử dụng.

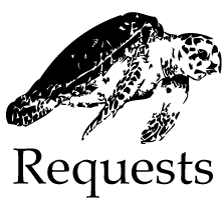
Khác với các ngôn ngữ khác, khối lệnh của Python không nằm trong cặp ngoặc nhọn như C, Java,… mà được thể hiện qua thụt các câu lệnh vào sâu hơn về phía phải. Điều này giúp người sử dụng dễ nhìn và nhận biết các khối lệnh hơn.



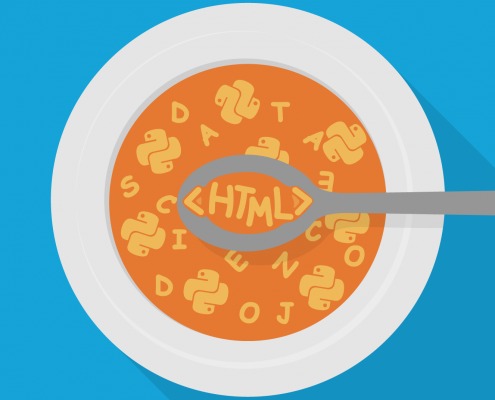
Khả năng mở rộng cũng là một lý do để bạn chọn Python. Nếu bạn biết về ngôn ngữ C, bạn có thể viết những dòng lệnh có tích hợp Python. Điều này được khá nhiều lập trình ứng dụng rộng rãi và một trong số đó là ứng dụng vào Arduino, vừa phát triển Arduino bằng C++ vừa phát triển giao diện hay tính toán bằng Python.

Các câu lệnh trong Python đều nằm trong các dòng mã ngồn. Bạn không cần phải nhớ bất kỳ kí tự nào khi kết thúc câu lệnh. Giống với hầu hết các ngôn ngữ khác, các cấu trúc điều khiển rẽ nhánh hay vòng lặp đều có ở Python (if, elif, else, for, while). Lập trình hướng đối tượng hay thiết kế module cũng là một lý do nhiều người lựa chọn Python để sử dụng. Python cho phép bạn phát triển module rồi sử dụng ở các chương trình khác. Python cũng cung cấp sẵn rất nhiều module chuẩn dược viết hỗn hợp bằng C và Python được gọi là CPython.

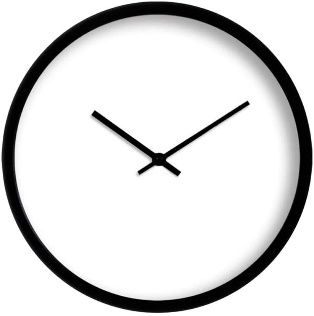
Để cài đặt các thư viện trong Python cực kì đơn giản với lệnh pip install tên\_thư\_viện và sau đó import tên\_thư\_viên là có thể sử dụng

Các thư viện được tôi sử dụng trong Web Crawler là BeautifulSoup, Requests, Os và Datetime

Đầu tiên là thư viện Requests. Module này cho phép chúng ta gửi yêu cầu HTTP và trả về tất cả dữ liệu của HTML bao gồm (mã hóa, nội dung, trạng thái,…) và phù hợp với hầu hết các phiên bản Python từ 2.x đến 3.x

Tại sao tôi sử dụng BeautifulSoup? Nếu bạn muốn lấy thông tin từ mọt trang web thì bạn cần dùng các API mà trang web đó cung cấp cho bạn. Cách này khác đơn giản nhưng không phải bạn đều sẽ được cung cấp API tất cả các trang web. Đây là thư viện cho phép ta trích xuất dữ liệu từ HTML tuy đơn giản mà hiệu quả. Với các module này bạn có thể trích xuất các thông tin như tiêu đề, văn bản, đường dẫn,… từ file code HTML đã được requests tải về.

Module Os cho phép chúng ta có thể thao tác với các tập tin và thư mục cũng như khai thác thông tin trong đó. Đây là module được cung cấp sẵn trong Python một cách linh động và phụ thuộc vào hệ điều hành. Thư viện này bạn chỉ cần import vào sử dụng và không cần cài đặt.

Thư viện này cũng giống với thư viện Os, nó được thêm vào sẵn và không cần cài đặt. Bạn có thể lấy dữ liệu ngày tháng và thời gian hiện tại thông qua module Datetime. Việc tôi chọn thư mục này sẽ giúp tôi có thể truy xuất lại trang web đã được thu thập dữ liệu vào ngày tháng và thời gian nào.

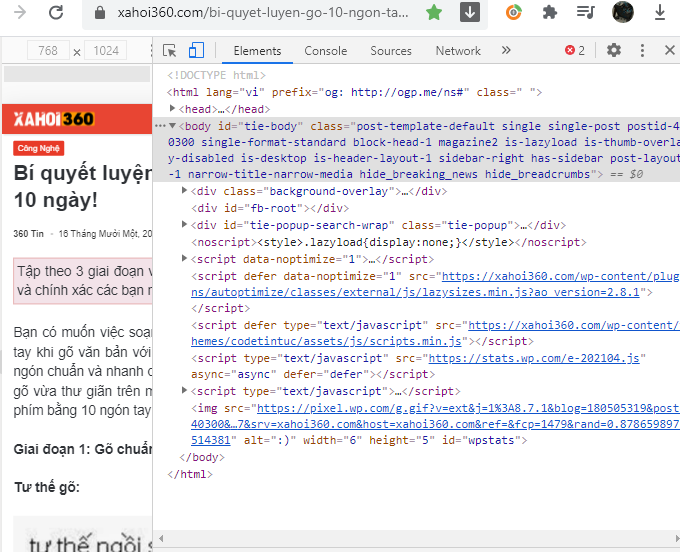
Với việc kết hợp cấu trúc đơn giản của Python và 4 thư viện trên, bạn hoàn toàn có thể làm một con bot thu thập dữ liệu một cách tự động hóa với một đường dẫn bất kì.

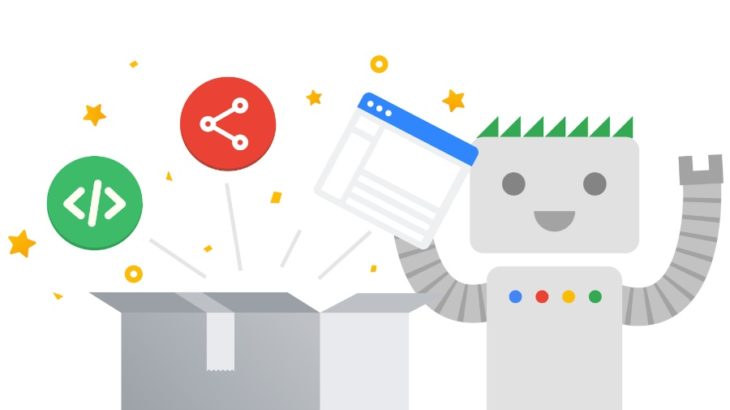
# **Chương 2 Thiết Kế Hệ Thống Web Crawler**

Trả lời cho những câu hỏi:

* Cơ chế hoạt động như thế nào ?
* Lưu đồ thuật toán ra sao ?
* Kết quả trả về là gì ?

## **2.1 – Kiến trúc hệ thống**

Mỗi thư viện đều có những chức năng riêng. Requests có thể yêu cầu truy cập và tải các trang web xuống. Thư viện BeautifulSoup có thể sử dụng trang web để lọc ra và cấu trúc lại thành file HTML. Thư viện Os giúp bạn lưu các dữ liệu vào một thư mục được chỉ định trong ổ cứng. Thư viện sẽ cho chúng ta thời gian để lưu trữ.

Nếu bạn muốn thu thập dữ liệu các trang web, bạn phải sử dụng các trình duyệt để truy cập vào trang web, bật Developer Tool để xem các dữ liệu từ file HTML. Sau đó lọc thông tin đường dẫn vừa cào dữ liệu, trích xuất ra các đường dẫn, ảnh, video,… rồi lưu lại vào thư mục. Việc làm thủ công này khác mất thời gian và công sức, dễ bị bỏ xót trang web.

Cơ chế hoạt động của Web Crawler giống như các bước bạn làm thủ công nhưng bây giờ máy sẽ tự động vận hành. Máy sẽ yêu cầu quyền truy cập vào trang web ban đầu, tải trang web về, cấu trúc và sắp xếp rồi xuất ra đoạn code HTML. Trong đoạn code HTML, máy bắt đầu lọc thẻ <a> và lấy tất cả các thông tin trong thẻ <a> rồi chọn lọc và chỉnh sửa để có những đường dẫn mới. Sau khi có tất cả thông tin cần thiết, máy sẽ lưu vào một thư mục được chỉ định. Với việc lặp đi lặp lại những thao tác này, bạn sẽ có 1 thư mục trong đó chứ tất cả dữ liệu đã thu thập được.

Tôi đã tạo ra 4 file python bao gồm main.py, web.py, folder.py và function.py. File main.py chính là kịch bản chính. Các dòng lệnh sẽ chạy theo thứ tự ở main.py. File web.py sẽ thực hiện các thao tác liên quan đến web như tải HTML, lọc đường dẫn, chỉnh sửa đường dẫn,…. File folder.py sẽ thực hiện các thao tác tạo và lưu trữ các tập tin, thư mục. File function.py sẽ thực hiện các thao tác làm sạch, chọn lọc và sắp xếp lại thông tin cần thiết.



Link code: <https://github.com/tranlybuu/Project_HUET_Web_Crawler>

## **2.2 – Lưu đồ thuật toán**

## **2.3 – Kết quả thu được**

Tạo danh sách chứa url chưa duyệt

Tạo danh sách chứ các url đã duyệt

Tạo biến count = đếm số url đã duyệt

Lưu url vào tập tin chứ url đã cào

Tạo tập tin số 3 chứa các url vừa tìm thấy

Tạo tập tin số 2 chứa code HTML

Tạo tập tin số 1 chứa url vừa cào và thời gian cào

Di chuyển vào thư mục vừa tạo

Lưu url vào tập tin chứ url lỗi

Lưu vào danh sách url lỗi

Lỗi

Tải HTML từ url

Lấy ra 1 url trong danh sách chưa duyệt

Tạo thư mục đặt tên tự động

Lưu vào danh sách url đã cào

Thêm vào danh sách url chưa duyệt

Lọc ra những đường dẫn chưa được tìm thấy trước đây

Chỉnh sửa các đường dẫn còn thiếu

Lấy toàn bộ đường dẫn

Lọc các đường dẫn hợp lệ

Kết thúc chương trình

Sai

Biến count nhỏ hơn giới hạn url tải và danh sách chờ lớn hơn 0

Đúng

Kiểm tra ổ cứng C:\

Tạo thư mục CRAWLER

Tạo tập tin chứ url lỗi

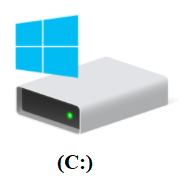
Tạo tập tin chứ url đã duyết

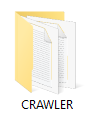
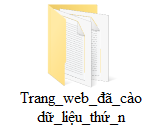
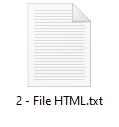
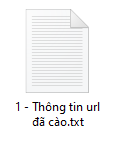
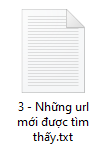
Sau khi Bot hoạt động, bạn sẽ thu được một thư mục tại ổ cứng C:\ là một thư mục CRAWLER. Trong thư mục CRAWLER sẽ có 3 thành phần chính.

Đầu tiên, History.txt là tập tin chứ các đường dẫn đã được duyệt qua và đã thu thập được thông tin cần thiết.

Tiếp theo là tập tin Error.txt. Đây là tập tin chứ các đường dẫn bị lỗi (không thể truy cập vào tên miền).

Cuối cùng là các thư mục chứ nội dung các trang web đã tải về. Mỗi thư mục chứa nội dung trang web đã thu thập được cũng có 3 tập tin. Tập tin số 1 là chứa ngày/ tháng/ năm, thời gian và đường dẫn vừa cào dữ liệu. Tập tin số 2 chính là file code HTML. Và tập tin thứ 3 chứa tất cả các đường dẫn vừa tìm được.





# **Chương 3 Kết Luận Về Đồ Án**

Trả lời cho những câu hỏi:

* Ưu điểm
* Nhược điểm
* Tài liệu kham khảo

**Đánh giá hệ thống: 80%**

Độ chính xác: 95% (Em thấy phần lọc các đường dẫn chưa tối ưu)

Chưa lọc được hình ảnh, video, từ khóa

***Ưu điểm:***

Tự động hóa với tốc độ cao các thao tác thủ công giúp giảm tải số lượng công việc, thao tác.

Dễ dàng so sánh các trang web giúp ta có thể phân tích thị trường

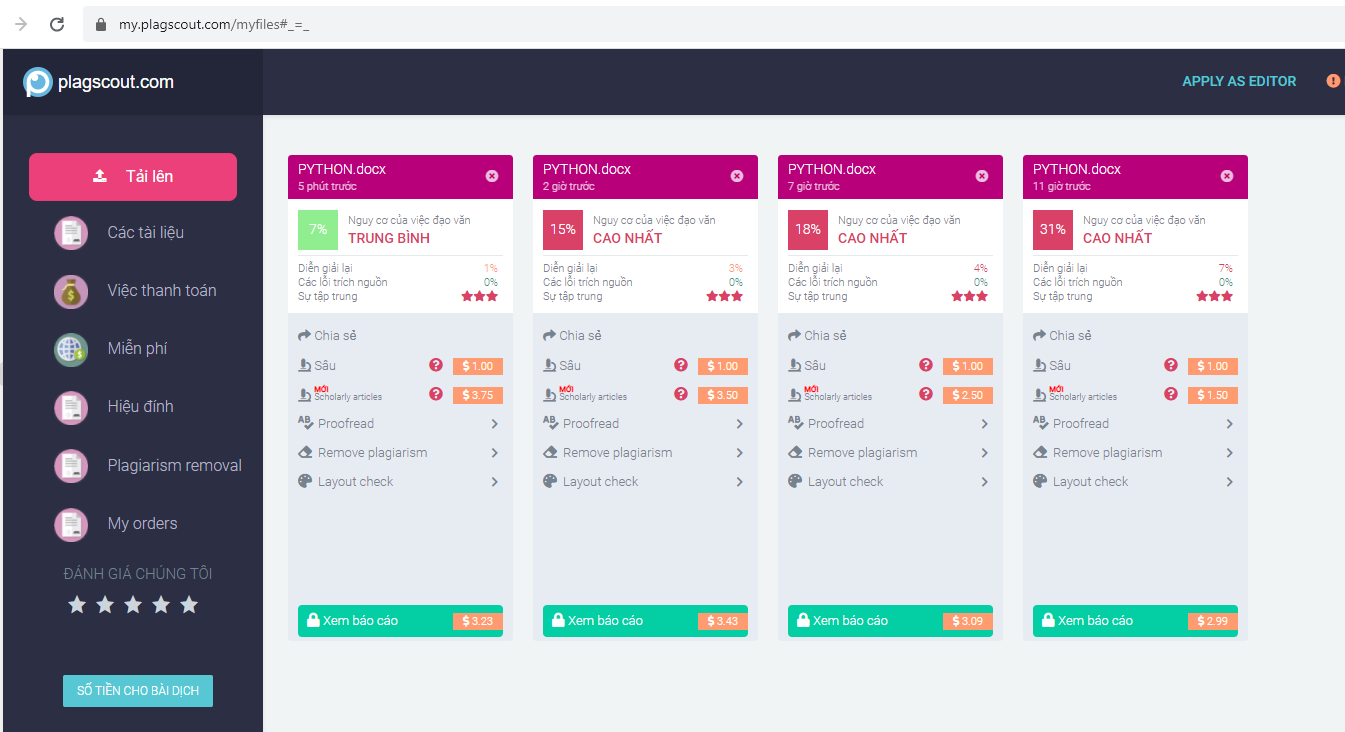
Tăng lượt view của trang web giúp trang web có khả năng được đề xuất, mọi người biết được nhiều hơn về trang web

***Nhược điểm:***

Phải điều chỉnh lại khá nhiều nếu cấu trúc html của website bị thay đổi

Có thể xảy ra thiếu xót trong lúc lọc các đường dẫn

**Đánh giá của bản thân:** Lần đầu làm đồ án đối với em cũng không quá khó khăn để thực hiện. Em đã nhận ra trước đây mình code khá lộn xộn và mỗi lần đọc lại mất khá nhiều thời gian. Chương trình vẫn tối ưu nhất vì chưa thể lọc ra các hình ảnh và video. Em nhận được nhiều kiến thức mới sau đồ án này và sẽ cải thiện cho những bài sau. Em cảm ơn sự giúp đỡ của thầy.

**Tỉ lệ 7% đạo văn**

**Tài liệu kham khảo:**

# Bibliography

GPSC, G. (2021, 1 31). *GiaiPhapSEO GPSC*. Được truy lục từ GiaiPhapSEO GPSC: https://giaiphapseo.com/web-crawler-la-gi/

Wikipedia. (2021, 1 6). *HTML*. Được truy lục từ Wikipedia: https://vi.wikipedia.org/wiki/HTML

Wikipedia. (2021, 1 5). *Python*. Được truy lục từ Wikipedia: https://vi.wikipedia.org/wiki/Python\_(ng%C3%B4n\_ng%E1%BB%AF\_l%E1%BA%ADp\_tr%C3%ACnh)