

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO**

**MÔN THƯƠNG MẠI ĐIỆN TỬ**

**Đề tài: Tạo website so sánh giá với web crawler**

Giảng viên hướng dẫn: Thầy Văn Đức Sơn Hà

Lớp: IS334.J21

Nhóm sinh viên thực hiện:

1. Bùi Thị Huyền Trân. MSSV: 16521275

2. Nguyễn Thị Kim Yến. MSSV: 16521485

3. Bùi Nguyên Mão. MSSV: 16520724

TPHCM, tháng 5, năm 2019

Mục lục

[1 Công nghệ web crawler và ứng dụng 3](#_Toc9890057)

[1.1 Công nghệ web crawler 3](#_Toc9890058)

[a) Giới thiệu 3](#_Toc9890059)

[b) Các hướng ứng dụng chính của web crawler 3](#_Toc9890060)

[1.2 Ứng dụng 4](#_Toc9890061)

[a) GoogleBot 4](#_Toc9890062)

[b) Xenon 4](#_Toc9890063)

[2 Dịch vụ so sánh giá 4](#_Toc9890064)

[2.1 Giới thiệu về dịch vụ so sánh giá 4](#_Toc9890065)

[2.2 Công nghệ sử dụng 5](#_Toc9890066)

[2.3 Các dịch vụ so sánh giá nổi tiếng trên thế giới: 6](#_Toc9890067)

[3 Các sản phẩm so sánh giá thị trường Việt Nam: 6](#_Toc9890068)

[4 Nội dung đồ án: 7](#_Toc9890069)

[4.1. Giới thiệu: 7](#_Toc9890070)

[4.2 Hiện thực sản phẩm: 7](#_Toc9890071)

[5. Kết luận - đánh giá 11](#_Toc9890072)

[6. Hướng phát triển 11](#_Toc9890073)

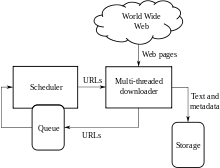
[7 Nguồn tham khảo: 11](#_Toc9890074)

# Công nghệ web crawler và ứng dụng

* 1. **Công nghệ web crawler**

1. **Giới thiệu**

* Trình thu thập web (web crawler) là một chương trình khai thác dữ liệu thông qua cấu trúc html của một trang web để phục vụ nhiều mục đích khác nhau.
* Trong thời kỳ đầu, công nghệ này có nhiều tên gọi như web spider, nhưng ngày nay tên gọi phổ biến nhất vẫn là trình thu nhập web. Mặc dù vậy cụm từ ‘thu thập’ không lột tả được hết tốc độ của những chương trình này, vì chúng có tốc độ làm việc đáng kinh ngạc, có thể thu thập dữ liệu lên đến hàng chục ngàn trang trong vòng một vài phút.
* Từ thời kỳ đầu, một động lực quan trọng thúc đẩy quá trình phát triển của việc thiết kế trình thu thập web là lấy được nội dung các trang web và thêm chúng hoặc đường dẫn của chúng vào một kho lưu trữ trang – một kiểu kho lưu trữ có thể dùng để phục vụ cho các ứng dụng cụ thể trong công cụ tìm kiếm web (search engine).
* Về bản chất, quá trình thu thập web chính là quá trình duyệt đệ quy một đồ thị. Các web được xem như một đồ thị với các trang là các đỉnh (node) và các siêu liên kết là các cạnh. Quá trình lấy trang và trích xuất các liên kết bên trong nó tương tự như việc mở rộng tìm kiếm một đỉnh trong đồ thị. Việc tìm kiếm này là khác nhau trong các trình thu thập sử dụng chiến lược tìm kiếm khác nhau.
* Các Trang web chủ yếu được viết bằng các ngôn ngữ đánh dấu văn bản như HTML và XHTML. Web crawler là kỹ thuật bóc tách và xử lý đoạn mã đã được cấu trúc để thu lại các khối dữ liệu cần thiết. Tốc độ làm việc của một ứng dụng web crawler diễn ra rất nhanh, có thể thu thập dữ liệu hàng ngàn trang chỉ trong vài phút.



1. **Các hướng ứng dụng chính của web crawler**

* Thu thập tất cả đường dẫn trên trang web phục vụ các công cụ tìm kiếm: Các web crawler thường bắt đầu bằng cách sử dụng các đường dẫn ứng với trang đầu tiên của website, đọc các nội dung trang web và tìm ra tất cả các đường dẫn siêu văn bản của trang, rồi tiếp tục thu thập từ các đường dẫn này cho đến trang cuối cùng, quá trình này sẽ thu về tất cả dữ liệu đường dẫn của một website và lưu trữ lại với chỉ mục theo các thuật toán sắp xếp độ ưu tiên.
* Thu thập phần nội dung, tài nguyên định trước phù hợp yêu cầu: Ứng dụng cần phải được xác định sẵn khu vực thu thập dữ liệu cho từng websites và cần phải có bộ đường dẫn mục tiêu của quá trình khai thác. Ứn dụng vào các website thu thập dữ liệu, so sánh giá,...

Các dữ liệu được hiển thị trên các trang web là các thông tin công khai và có thể dễ dàng được sao chép, thu thập một cách thủ công. Nhưng với các công cụ thu thập các request được gửi đi nhiều, nhanh tùy thuộc vào công cụ sẽ gây ra một lượng lớn xử lý ở phía server có thể gây ra tình trạng quá tải. Vì vậy, có một số biện pháp để giảm thiểu tình trạng bị crawl dữ liệu:

* Chặn các truy cập từ tất cả các ip có hành động gửi quá nhiều request đến server hoặc có biện pháp đặt thêm các rào chắn truy cập như capcha.
* Có thể ngăn chặn hoặc giảm thiểu tần suất crawl của GoogleBot trong quá trình index site với google search console.
  1. **Ứng dụng**

1. **GoogleBot**

* Là phần mềm web crawler của google được dùng để thu thập thông tin các trang để xây dựng, cập nhật chỉ mục có thể tìm kiếm cho google search engine.
* Chủ yếu thu thập dữ liệu theo các thuộc tính href và src của thẻ html nhưng cũng có nhiều nguồn tin nói rằng GoogleBot có thể thu thập dữ liệu với javascript và phân tích nội dung được gọi với ajax.[2]

Link: <https://search.google.com/search-console/welcome>

**b) Xenon**

- Là phần mềm tìm kiếm và bí mật giám sát internet, hiện đang được công khai sử dụng bởi các cơ quan thuế của 6 quốc gia: Hà Lan, Áo, Canada, Đanh Mạch, Anh, Thụy Điển để điều tra các khả năng trốn thuế của các trang web (Cửa hàng trực tuyến, trang đánh bạc,…) và các khách hàng bán hàng, đấu giá trực tuyến. Phần mềm sử dụng công nghệ web crawler để thu thập dữ liệu. Có một số ý kiến trái chiều về quyền tự do cá nhân của phần mềm trên.

Link: <https://en.wikipedia.org/wiki/Xenon_(program)>

# Dịch vụ so sánh giá

**2.1 Giới thiệu về dịch vụ so sánh giá**

**-** Công cụ so sánh giá, hoặc websites so sánh giá là một công cụ tìm kiếm mà người mua hàng sử dụng để tìm kiếm, lọc và so sánh các sản phẩm dựa trên giá, tính năng, đánh giá và các tiêu chí khác để đưa ra quyết định mua hàng.

- Hầu hết các trang so sánh giá không thu tiền từ người sử dụng và cũng không trực tiếp bán sản phẩm mà chỉ tổng hợp danh sách sản phẩm từ nhiều nhà bán lẻ khác nhau, doanh thu sẽ đến từ các thỏa thuận tiếp thị liên kết. Tùy vào mô hình cụ thể, các nhà bán lẻ có thể cần trả phí để được hiển thị trên trang web hoặc với mỗi lượt nhấp chuột của khách hàng.

- BargainFinder được cho là công cụ so sánh giá đầu tiên được phát triển vào năm 1995 bởi Andersen Consulting. Tiếp sau đó, nhiều dịch vụ so sánh giá đã xuất hiện và phát triển lớn mạnh như Excite, Junglee( được Amazon mua lại), NexTag,….Đến khoảng năm 2010, các trang web so sánh giá bắt đầu phát triển ở thị trường Đông Nam Á với SoXpress tại Singapore, và trong những năm tiếp theo Baoxian( Trung Quốc), Jirnexu (Malaysia)…[4]

- Vào năm 2017, google bị phạt 2,7 tỷ usd vì đã có hành vi lạm dụng công cụ tìm kiếm để quảng bá dịch vụ google shopping ở đầu kết quả tìm kiếm, giảm lượt truy cập đối với các đối thủ cạnh tranh.[6]

- Theo thống kê tại vương quốc Anh, có hơn 70% số lượng người dùng internet truy cập vào các trang web so sánh giá mua sắm [5]. Bốn dịch vụ so sánh giá lớn nhất tại Anh đã tạo ra doanh thu (1,2 tỷ usd) trong năm 2013 và lợi nhuận trung bình năm của nhóm tăng lên 14% trong năm đó.[5] Có rất nhiều sự đóng góp của so sánh giá trong việc mở rộng ngành công nghiệp thương mại điện tử.

- Ban đầu, các dịch vụ so sánh tách rời nhau các tính năng so sánh giá sản phẩm, đánh giá nhà cung cấp và đánh giá sản phẩm nhưng sau đó dần hợp nhất lại thành các công ty dịch vụ so sánh.[7]

* 1. **Công nghệ sử dụng**

Các công nghệ được sử dụng để tạo trang web so sánh giá:

* Các trang so sánh giá nhận dữ liệu từ các nhà phân phối sản phẩm thông qua dữ liệu danh sách sản phẩm và giá dạng liệt kê để trang so so sánh giá nhập vào cơ sở dữ liệu hoặc bằng các API được cung cấp từ đối tác. Dữ liệu và giá của sản phẩm cần được lưu trữ, cập nhật, chỉnh sửa theo thời gian nên dữ liệu có thể có sai sót khi không được cung cấp kịp thời.
* Các trang so sánh giá xây dựng phương pháp thu thập dữ liệu bằng web crawler. Phương pháp này chủ yếu được dùng với các trang web độc lập và nhỏ hơn. Dữ liệu được cập nhật real time theo giá niêm yết và không cần tốn bộ nhớ để lưu trữ.

- Trang so sánh giá thu thập dữ liệu thông qua mô hình crowdsourcing, từ những người dùng đóng góp, đánh giá như một mạng xã hội hay diễn đàn thảo luận.

* 1. **Các dịch vụ so sánh giá nổi tiếng trên thế giới:**
* NexTag: công ty Mỹ cung cấp dịch vụ so sánh giá độc lập cho các sản phẩm du lịch và dịch vụ, được thành lập vào 1999. năm 2008, NexTag được xếp vào 50 trang web tốt nhất.[8]  Sau khi có nhiều động thái mua các công ty thương mại, nền tảng mạng xã hội vào năm 2011, đã biến NexTag trở thành cửa hàng thương mại điện từ đa dạng sản phẩm. Ngày 28 – 8 – 2018, NexTag dừng hoạt động.
* Google Shopping: là một dịch vụ của google phát hành vào 2002 với tên gọi ban đầu là Froogle. Ban đầu dịch vụ theo mô hình liệt kê giá được cung cấp bởi các nhà phân phối và kiếm tiền thông qua Adwords nhưng từ 2012, dịch vụ trở thành mô hình trả phí nơi nhà phân phối trả tiền để được liệt kê sản phẩm.
* Trivago: là công cụ so sánh giá khách sạn hàng đầu thế giới, hỗ trợ so sánh hơn 400 websites với hơn 1,8 triệu khách sạn ở hơn 190 quốc gia.

# Các sản phẩm so sánh giá thị trường Việt Nam:

* Ở Việt Nam một vài năm trở lại đây, thương mại điện tử bắt đầu bùng nổ kéo theo sự xuất hiện của các dịch vụ so sánh giá. các trang web này thường sẽ thống kê dữ liệu từ các nhà bán hàng và đưa ra sắp xếp, gợi ý cho khách hàng mà không trực tiếp bán sản phẩm. Doanh thu sẽ đến từ các thỏa thuận chiết khấu theo số lượng giới thiệu.
* Hiện nay đã có 1 số trang web có tính năng tương tự phục vụ tìm kiếm giá ở nhiều cửa hàng với nhiều loại sản phẩm. Theo đánh giá chung, các trang web hiện tại đang có những điểm sau:
* Giao diện người dùng được cập nhật hiện đại dễ sử dụng.
* Giá sản phẩm được cập nhật từ nhiều nguồn với đa dạng các loại mặt hàng, độ chính xác cao.
* Tuy nhiên, cũng có những hạn chế:
* Tuy các trang web có nhiều loại mặt hàng nhưng vẫn còn chia ra các nhóm ngành ví dụ như: nhóm website so sánh hàng tiêu dùng, so sánh giá dịch vụ vận tải, du lịch, dịch vụ bảo hiểm,…
* Các website hiện tại thường được cập nhật giá dựa trên dữ liệu từ các gian hàng nên dữ liệu giá có thể sẽ được cập nhật chậm hơn và cần có cơ sở dữ liệu để lưu giá sản phẩm trong quá trình hoạt động
* Các yếu tố về chất lượng sản phẩm hoặc độ uy tín của gian hàng cũng chưa được đảm bảo.
* Thị trường ở Việt Nam hiện tại đang hướng đến chủ yếu các mặt hàng công nghệ, thời trang hoặc một số nhỏ về các sản phẩm dịch vụ vé máy bay, dịch vụ đi lại và một số rất ít về so sánh các dịch vụ bảo hiểm, tín dụng. Nhu cầu người dùng về các công cụ phân tích giá là không hề thiếu và còn rất nhiều ngành hàng chưa có công cụ đánh giá, so sánh như về các dịch vụ giáo dục (so sánh chất lượng, giá các trung tâm tiếng Anh, các cơ sở giáo dục Đại Học,...), y tế (giá, chất lượng các dịch vụ chăm sóc sức khỏe,...),...

# Nội dung đồ án:

## Giới thiệu:

1. Lý do chọn đề tài:

* Do nhu cầu thu thập thông tin của con người ngày càng tăng, lượng thông tin trên internet rất phong phú nên vấn đề tổng hợp thông tin ngày càng trở nên bức thiết. Với một lượng dữ liệu lớn việc thu thập bằng tay tốn rất nhiều công sức, và không đạt hiệu quả cao, chính vì thế cần một công nghệ có thể tổng hợp thông tin một cách tự động và trình thu thập web đã ra đời. Đề tài đặt ra vấn đề tìm hiểu về trình thu thập thông tin trên web và bước đầu sẽ xây dựng một trang web có khả năng tổng hợp thông tin tự động từ một số website bán hàng như https://fptshop.com.vn , <https://thegioididong.com> và https://vienthonga.vn/ . Từ đó so sánh giá và đưa ra những gợi ý mua hàng cho người dùng.

1. Mục tiêu

* Tìm hiểu công nghệ web crawler và các ứng dụng của nó trong cuộc sống hiện đại, có cái nhìn tổng quát về so sánh giá và có đánh giá, chiến lược đúng đắn cho thị trường này.
* Sản phẩm: xây dựng thành công website so sánh giá sản phẩm công nghệ đơn giản với công nghệ web crawler, truy cập website của các nhà phân phối, phân tích cấu trúc html và lấy thông tin về giá cũng như 1 số đặc tính khác của sản phẩm rồi làm dữ liệu để hiện lên trang. Dữ liệu được thu thập từ 3 trang: thegioididong.com, fptshop.vn, vienthonga.com

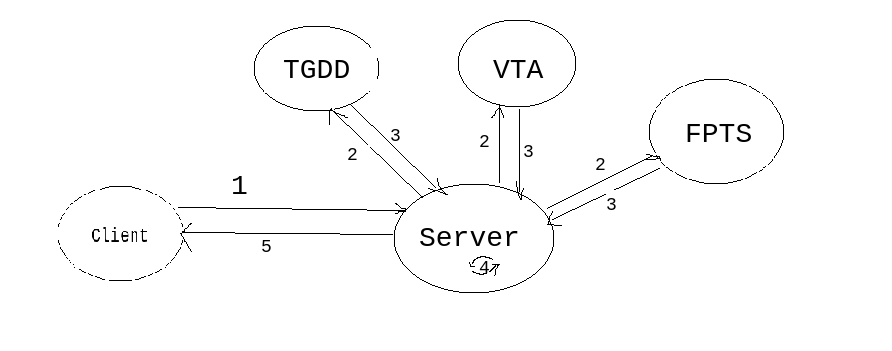
1. Tầm nhìn

* Có thể rút ra kết luận và những đánh giá khách quan về sản phẩm đã tạo ra, đưa ra những biện pháp tối ưu và phát triển cho ứng dụng.

## Hiện thực sản phẩm:

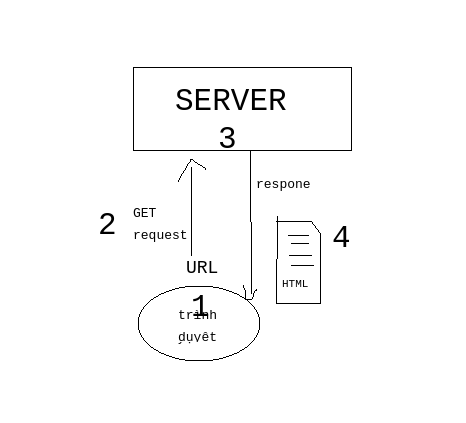
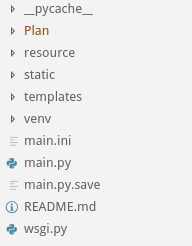
1. Mô hình xây dựng

workflow của website:

* Người dùng gửi yêu cầu là tên sản phẩm nhập từ ô tìm kiếm rồi gửi lên cho server xử lý.
* Server sẽ tiến hành đi crawl dữ liệu từ các trang nguồn với keyword mà client yêu cầu.
* Từ các dữ liệu nhận được từ các trang nguồn, server sẽ tiến hành kiểm tra các sản phẩm rồi sắp xếp theo thứ tự giá thấp đến cao.
* Server sẽ trả lại client respone gồm danh sách các sản phẩm tìm thấy từ keyword và render thành giao diện phù hợp.
* Người dùng sẽ xem thông tin sản phẩm phẩm, quyết định và sẽ được dẫn đến nơi bán.
* Đặc điểm:
* Dữ liệu giá, thông tin sản phẩm được crawl và xử lý ngay khi người dùng bấm nút tìm kiếm nên hoàn toàn chính xác theo thời gian thực.
* Các xử lý, dữ liệu chỉ cần gửi đi một lần mà không cần phải lưu trữ với cơ sở dữ liệu.
* Ngôn ngữ lựa chọn:
* Sử dụng python + flask framework làm ngôn ngữ xử lý phần backend, jinja2 template engine + html, css, js để xây dựng frontend.

1. Xây dựng website

* Python là ngôn ngữ lập trình scripting hiện đại với cú pháp ngắn gọn, đơn giản, gần với ngôn ngữ tự nhiên nên rất dễ tìm hiểu. Bên cạnh đó, có rất nhiều bộ thư viện được xây dựng với python biến nó trở thành một ngôn ngữ đa năng.
* Flask là một micro-web-framework viết trên ngôn ngữ python có ưu điểm nhẹ dễ dàng phát hiện và xử lý các lỗi bảo mật. Flask bao gồm jinja2 template engine giúp việc viết các file html rõ ràng hơn.
* Cấu trúc một dự án web với flask:



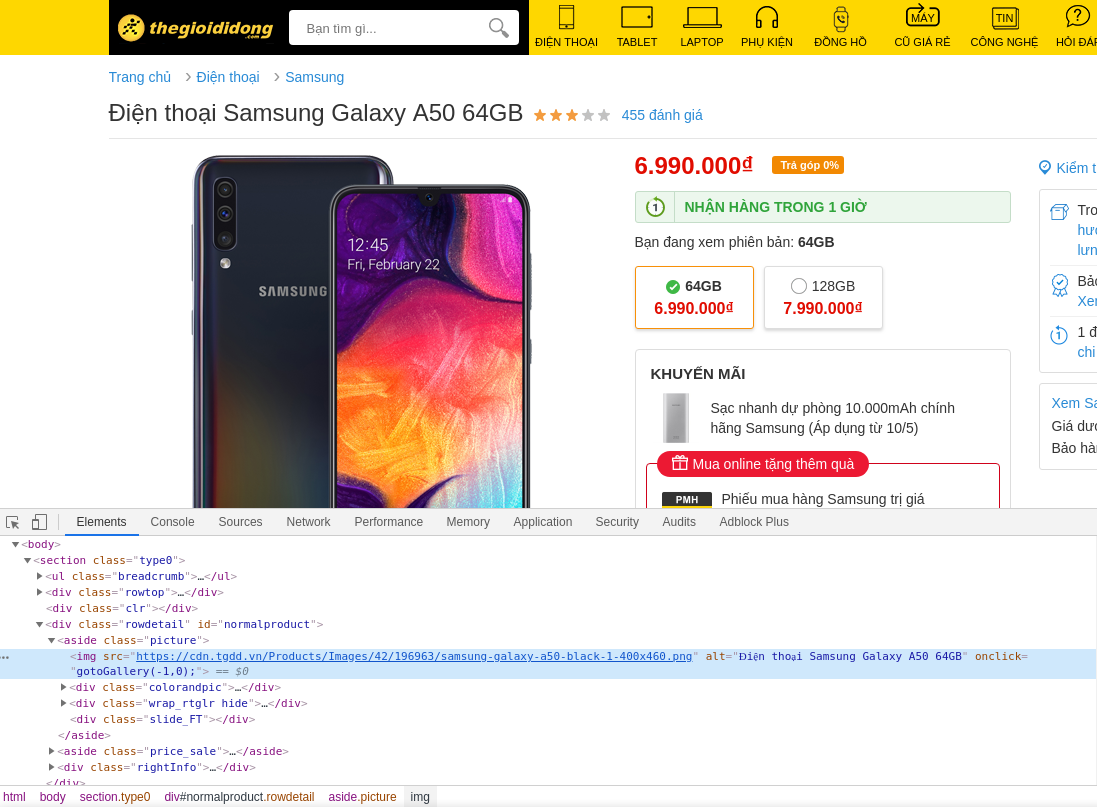
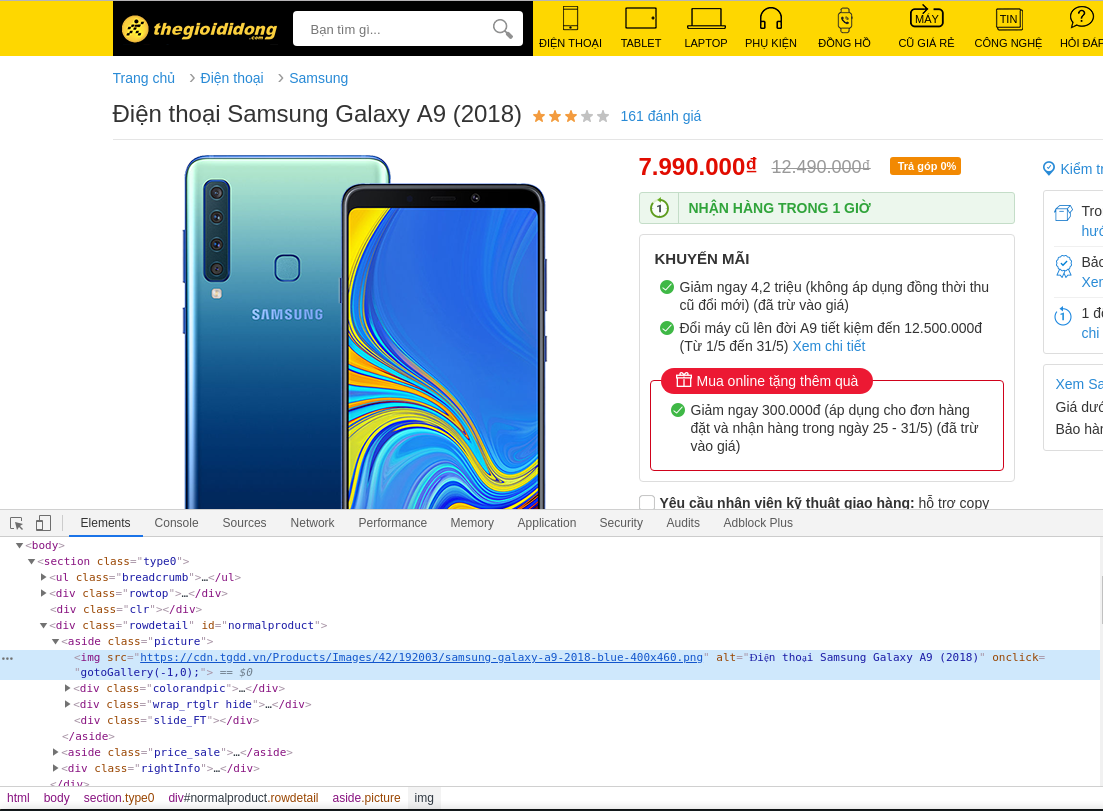
* Statics là nơi chứa các file images, css, java script
* templates chứa các file giao diện html
* main.py là file dựng server. Ở đây

/// đợi code xong sẽ thêm mấy phần file code vào đây

///Hình ảnh các giao diện cũng sẽ được cho vào đây.

1. Xây dựng module crawl dữ liệu từ các trang bán hàng

* Nhiệm vụ của module là chứa các xử lý liên quan việc thu thập dữ liệu theo từ khóa với các trang nguồn, sắp xếp thứ tự ưu tiên nơi bán theo giá sản phẩm.
* Kỹ thuật thu thập dữ liệu từ websites:
* Một websites được định danh với một url, khi truy cập với url đó, có nghĩa là máy tính đã gửi đến server đích của url một get request yêu cầu dữ liệu để hiển thị giao diện, lúc này server sẽ xử lý và trả về một file html chứa định dạng trang web
* Dữ liệu trang web được truyền tải dưới dạng ngôn ngữ đánh dấu có cấu trúc



* Trong một websites các trang có cấu trúc html giống nhau, chỉ có nội dung là thay đổi vì vậy để thu thập được nội dung của một trang, ta cần rút ra các quy luật bao gồm các id, class để chọn được đúng vị trí mà nội dung xuất hiện.
* Xây dựng module với python:
  + Thư viện requests có tác dụng tạo các get request đến các url chỉ định, thư viện bs4 có các hàm liên quan đến xử lý các dữ liệu với file html.
  + Trước tiên cần có danh sách các url chứa nội dung cần crawl. các trang nguồn có url riêng phục vụ mục đích tìm với keyword:

/// Có hình hàm tìm kiếm của các trang nguồn

* Với các trang danh sách sản phẩm tìm kiếm được ta đã có thể tạo ra được một bộ dữ liệu kiểu dictionary cho trang danh sách sản phẩm tìm kiếm.

/// Hàm danh sách sản phẩm

* Sau khi chon sản phẩm, ta cần lấy ra dữ liệu chi tiết cho sản phẩm ứng với trang nguồn gợi ý

/// Hàm lấy dữ liệu chi tiết sản phẩm

# Kết luận - đánh giá

* Tổng kết lại, thông qua đồ án, nhóm đã tìm hiểu được phần nào về thị trường công cụ so sánh giá, công nghệ web crawl và hoàn thành một website so sánh đơn giản quy mô nhỏ.
* Ưu điểm:
* Hạn chế:

# Hướng phát triển

* Như đã đánh giá ở trên, thị trường đang có rất lớn và thương mại điện tử Việt Nam đang ngày càng phát triển nên công cụ có rất nhiều hướng để phát triển.
* Thay đổi ngành hàng hoặc xây dựng với quy mô mặt hàng rộng lớn cả sản phẩm lẫn dịch vụ, từ cả y tế đến giáo dục.
* Đẩy mạnh các chương trình liên kết với các gian hàng, đảm bảo chất lượng sản phẩm để củng cố lòng tin người dùng và xây dựng thương hiệu như một kênh giới thiệu, đánh giá uy tín, trách nhiệm.
* Tích hợp thu thập thông tin người dùng, thói quen mua hàng để xây dựng tính năng gợi ý sản phẩm dựa trên công nghệ ML và AI.

# Nguồn tham khảo:

[1]https://en.wikipedia.org/wiki/Web\_crawler

[2]https://en.wikipedia.org/wiki/Googlebot

[3]<https://www.ukprwire.com/Detailed/Computers_Internet/Shopping_Comparison_Engines_market_worth_120m-_140m_in_2005_says_E-consultancy_1648.shtml>

[4]<https://www.postonline.co.uk/2319717/the-rise-of-price-comparison-sites-in-south-east-asia>

[5]<https://web.archive.org/web/20160912193044/http://www.phonesandyou.co.uk/blog/price-comparison-shopping-in-uk/>

[6]<https://www.bbc.com/news/technology-40406542>

[7]<https://www.researchgate.net/publication/265955076_A_Classification_of_Product_Comparison_Agents>

[8]<http://content.time.com/time/specials/2007/article/0,28804,1809858_1809955_1811450,00.html>

[9]<https://www.ecommercebytes.com/2019/01/07/what-happened-to-comparison-shopping-engine-nextag/>

[10]<http://flask.pocoo.org/>