

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

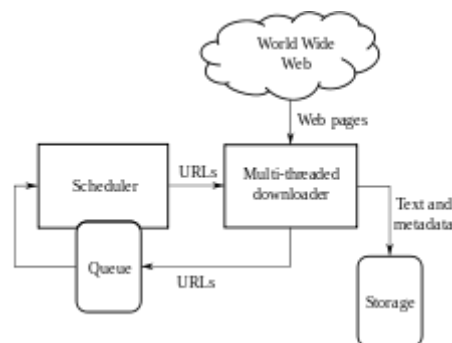
1	Công nghệ web crawler và ứng dụng	3
1.1	Công nghệ web crawler	3
a)	Giới thiệu	3
b)	Các hướng ứng dụng chính của web crawler	3
c)	Lưu ý khi crawler dữ liệu	4
1.2	Ứng dụng	5
a)	GoogleBot	5
b)	Xenon	6
2	Dịch vụ so sánh giá	7
2.1	Giới thiệu về dịch vụ so sánh giá	7
2.2	Công nghệ sử dụng	8
2.3	Các dịch vụ so sánh giá nổi tiếng trên thế giới:	8
2.4	Các sản phẩm so sánh giá thị trường Việt Nam:	12
2.5	So sánh một số website so sánh giá ở Việt Nam	13
2.6	Lợi ích của web so sánh giá	17
3	Nội dung đồ án:	18
3.1.	Giới thiệu:	18
a)	Lý do chọn đề tài:	18
b)	Mục tiêu	18
c)	Tầm nhìn	18
3.2	Hiện thực sản phẩm:	18
a)	Mô hình xây dựng	18
b)	Xây dựng website	19
c)	Xây dựng module crawl dữ liệu từ các trang bán hàng	21
4.	Sản phẩm - đánh giá	22
5.	Hướng phát triển	23
6.	Nguồn tham khảo:	24

1 Công nghệ web crawler và ứng dụng

1.1 Công nghệ web crawler

a) 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 như làm dữ liệu phân tích, đăng tải lên website, chia sẻ lên data center...
- 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.



b) 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. Ứng 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 captcha.
- Có thể ngăn chặn hoặc giảm thiểu tần suất crawl của GoogleBot trong quá trình index site với google search console.

c) Lưu ý khi crawler dữ liệu

- Nếu bạn lấy tin tự động với mục đích phát triển website/ blog
 - Lấy những loại tin tức không/ ít vi phạm chính sách bản quyền từ các công cụ tìm kiếm: các bài thuốc dân gian, các hướng dẫn pha nước ép, các bài chia sẻ về công dụng của từng loại rau củ quả... đó là những nội dung mang tính cộng đồng.
 - Crawler của bạn phải đủ thông minh để tách toàn bộ dữ liệu thu được và tối ưu lại nó 1 cách tốt nhất có thể (1 phần mang lại nội dung khác biệt trong mắt các công cụ tìm kiếm). Đây là điều có thể làm được nhé 😊 chỉ là bạn làm nó ở mức nào thôi !.

- Nếu bạn lấy tin tự động với mục đích làm dữ liệu phân tích

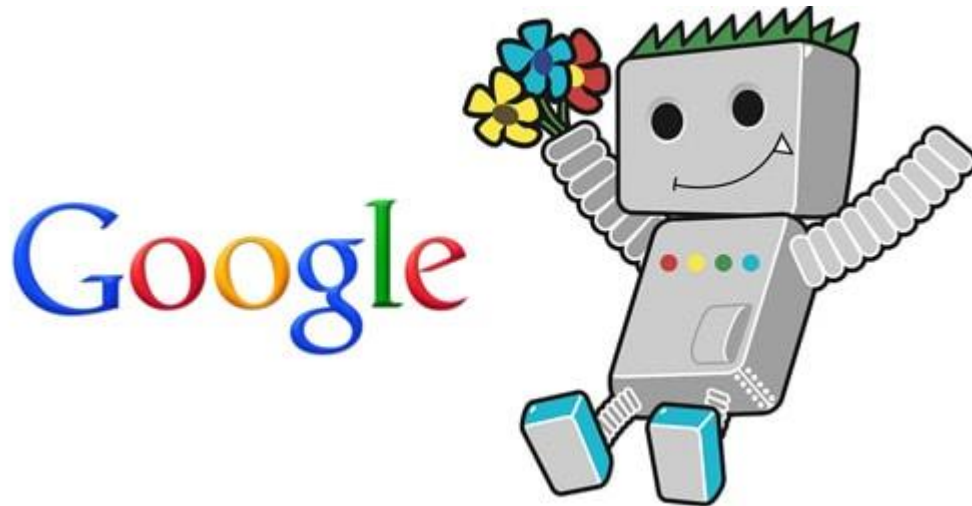
Trong trường hợp này dữ liệu bạn lấy về chỉ với mục đích phân tích nội dung đó thì bạn không cần những tiêu chí tối ưu cho các công cụ tìm kiếm (chắc chắn nó không cần thiết). Khi đó tiêu chí cơ bản như sau:

- Lấy bất cứ loại dữ liệu chứa thông tin từ bất cứ nguồn nào bạn muốn (sau này khi sử dụng nhớ trích nguồn ghi nhận công cho tác giả nhé :)).

- Crawler của bạn cần đủ thông minh để phân tích dữ liệu và thống kê theo tiêu chí của riêng bạn. Bạn đang bắt đầu làm công việc của Google tại nhà rồi đấy 😊

1.2 Ứng dụng

a) 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]



- Hoạt động: Googlebot sử dụng sơ đồ trang web và CSDL của các liên kết được phát hiện trong quá trình thu thập dữ liệu. Mỗi khi trình thu thập thông tin tìm thấy liên kết mới trên một trang web, nó sẽ thêm vào danh sách các trang để truy cập tiếp theo. Nếu Googlebot tìm thấy thay đổi trong các liên kết hoặc liên kết bị hỏng. Lúc này Googlebot sẽ nhận định chỉ mục có thể được cập nhật.
- Google không chia sẻ danh sách địa chỉ IP mà Googlebot sử dụng. Bởi vì những địa chỉ này thường xuyên thay đổi. Để biết Googlebot thực có ghé thăm trang web của bạn hay không, bạn thực hiện tìm kiếm IP ngược lại. Spammer hoặc kẻ mạo danh có thể giả mạo user-agent name, nhưng không phải địa chỉ IP. Bạn có thể tham khảo qua ví dụ xác minh tính hợp lệ của Googlebot.
- Bạn có thể sử dụng tệp robots.txt để xác định cách truy cập của Googlebot. Nhưng hãy thận trọng, nếu bạn làm sai, bạn có thể ngăn không cho Googlebot truy cập được. Điều này sẽ đưa trang web của bạn ra khỏi chỉ mục.
- Google Search Console:
 - Search Console là một trong những công cụ quan trọng nhất để kiểm tra khả năng thu thập thông tin của trang web. Nhờ đó có thể xác minh cách Googlebot nhìn thấy trang web của chúng ta.
 - Chúng ta cũng sẽ nhận được một danh sách các lỗi thu thập dữ liệu để khắc phục. Trong Search Console, ta cũng có thể yêu cầu Googlebot thu thập thông tin lại trang web của ta. Bên cạnh đó, để khắc phục những lỗi thu thập dữ liệu bạn cần kết nối Yoast SEO với Search Console, có thể nhập các lỗi và sửa chúng ngay từ phần phụ trợ của trang web. Yoast SEO Premium có thể làm nhiều hơn để làm SEO dễ dàng hơn.
- Tối ưu hóa với Googlebot : Bạn muốn Googlebot tối ưu thu thập dữ liệu trang web của bạn nhanh hơn? Đây là một thao tác kỹ thuật làm giảm không cho trình thu thập dữ liệu truy cập vào trang web của bạn một cách chính xác.
 - ⇒ Googlebot là 1 con robot truy cập trang Web của bạn. Web có những thao tác về kỹ thuật âm thanh thì Googlebot sẽ truy cập thường xuyên. Nếu bạn thường xuyên thêm nội dung mới, Googlebot thường xuyên xuất hiện.
 - ⇒ Đôi khi, bạn thực hiện các thay đổi quy mô lớn cho trang web của mình. Điều này bạn cần phải kiểm tra kỹ càng trình thu thập dữ liệu này cùng một lúc. Bởi vì những thay đổi có thể được phản ánh trong kết quả tìm kiếm.

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, Đan 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_\(program\)](https://en.wikipedia.org/wiki/Xenon_(program))

2 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, Jungle(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)...

- 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.

- 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. 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. 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

2.2 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 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.

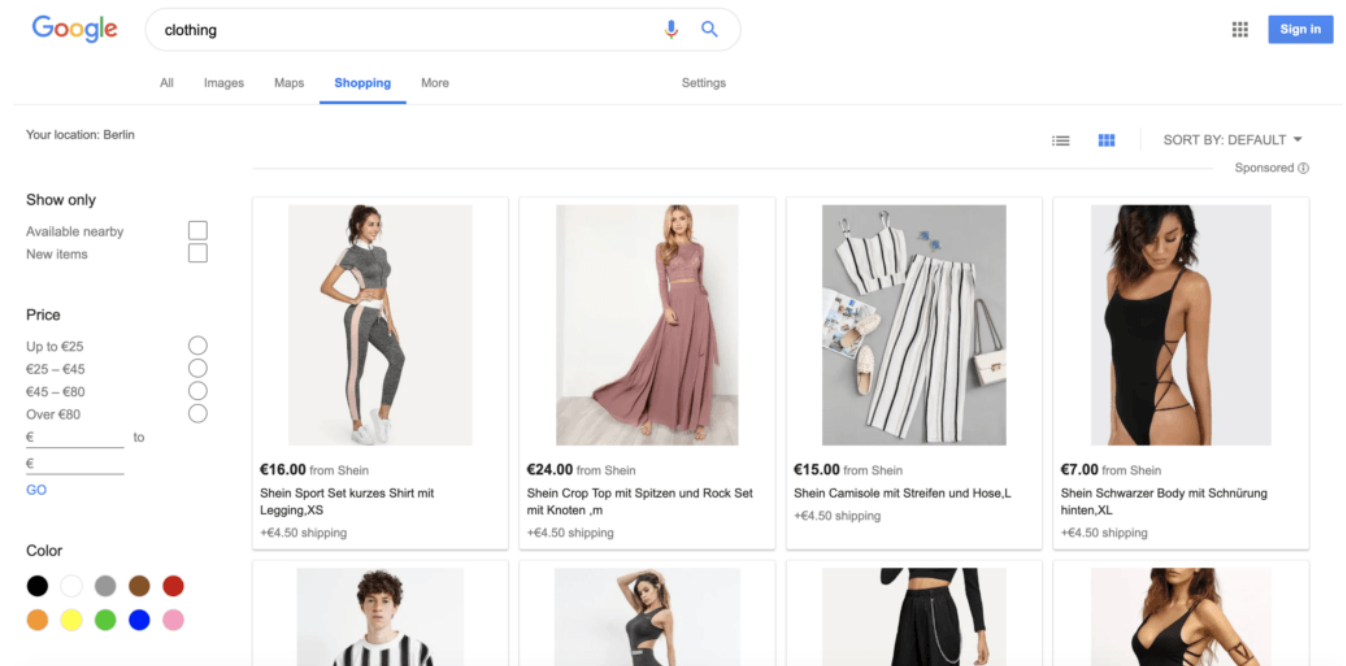
2.3 Các dịch vụ so sánh giá nổi tiếng trên thế giới:



Reviews and product comparisons

Next is a magazine that offers reviews, rankings and product comparisons on hundreds of everyday use items. Our team of writers compares, writes and advises on products ranging from computers to home appliances including DIY, sports or cars. Every 24 hours, our

- 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. 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 tử đa dạng sản phẩm. NextTag là một trong những công cụ mua sắm so sánh lâu đời nhất. Trang web so sánh giá này cho phép khách hàng đặt thông báo giá, so sánh người bán, tìm sản phẩm tương tự và hơn thế nữa. Khách hàng có thể tìm thấy các giao dịch tốt nhất từ các trang web như Amazon và eBay. Nhiều sản phẩm dropshipping nằm trên nền tảng của NextTag, cho phép người gửi hàng xem giá sản phẩm phổ biến.



- 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.
- Trang web so sánh giá của Yahoo Shopping tương tự như Google Shopping. Chủ cửa hàng có thể thêm sản phẩm của họ vào nền tảng. Bạn sẽ tìm thấy các sản phẩm được bán bởi các thương hiệu như Walmart và Amazon trên nền tảng.
- Công cụ mua sắm so sánh Yahoo khá dễ sử dụng. Bạn chỉ cần nhập sản phẩm bạn đang tìm kiếm vào thanh tìm kiếm và bạn sẽ được cung cấp các trang kết quả phù hợp với truy vấn tìm kiếm của bạn.

Home Mail Tumblr News Sports Finance Entertainment Lifestyle Answers Groups Mobile More

YAHOO! SHOPPING

Search Shopping Search Web Sign in Mail

Shopping Home
Clothing & Accessories
Electronics
Home & Garden
Flowers & Gifts
Toys & Baby
Computers
Movies & DVDs
Jewelry
Sporting Goods
More

Browse Shopping

Shop by Store | Shop by Brand | See all Categories

Clothing
Women's
Men's
Teen's
Shoes
Computers
Laptops
Desktops
Tablets
Printers

Electronics
Cameras
Cell Phones
Televisions
MP3 Players
Movies & DVDs
Action & Adventure
Kids & Family
Documentary
Foreign

Home & Garden
Appliances
Automotive
Bed & Bath
Furniture
Jewelry
Diamond Jewelry
Engagement
Watches
Jewelry

Flowers & Gifts
Flowers by Occasion
Flowers & Plants
Gifts by Occasion
Roses
Sporting Goods
Individual Sports
Exercise & Fitness
Camping & Outdoors
Fan Gear

Toys & Baby
Baby Girl's
Baby Boy's
Baby Gear
Nursery
More Categories
Books
Music
Health Care
Beauty

YAHOO! MAIL
Tự chọn màu yêu thích cho hộp thư của bạn.
Tải ứng dụng

Follow Yahoo Shopping

on Facebook on Twitter on Tumblr on Pinterest

Find, Compare, Read Reviews & Buy Online @ Yahoo Shopping - Online Shopping with great products, prices and reviews - Want to see your products in Yahoo

- BizRate cho phép khách hàng tìm giá tốt nhất, đặt thông báo giá và tìm kiếm thông qua vô số giao dịch trên công cụ tìm kiếm so sánh giá của họ. Nền tảng của BizRate cung cấp nhiều giao dịch sản phẩm tuyệt vời. Trang web thân thiện với người dùng và nhiều kết quả. Một vài tính năng khiến BizRate nổi bật bao gồm tùy chọn tải xuống các liên kết đến hướng dẫn sử dụng PDF cho hàng trăm thiết bị và tiện ích. Nó cũng có một tính năng cảnh báo giá, khá đơn giản để sử dụng. Tất cả những gì bạn phải làm là nhập địa chỉ email và ngưỡng giá và BizRate sẽ thông báo cho bạn bất cứ khi nào giá của sản phẩm bạn chọn nằm trong phạm vi cảnh báo của bạn.

bizrate.com search, compare, conquer.

GADGETS AND GEAR FOR THE GREAT OUTDOORS

It's time to get away from the computer and explore the great out of doors. Here's a roundup of some gear to use when you're camping, hiking, or just having fun outside.

READ MORE & SHOP OUR PICKS >>

1 2 3 4 5

WEEKLY DEALS

Save 30% Diamond Earrings

Save 50% Childrens Tablet

Save 50% Beats by Dre Pill

Save 65% Portable Heater

- Pronto có thể là một trang web so sánh giá hữu ích cho các chủ cửa hàng. Khi bạn tìm kiếm các sản phẩm trên Pronto, bạn sẽ tìm thấy nhiều trang web bán các sản phẩm tương tự như Walmart, Overstock, Bed Bath & Beyond, Amazon, v.v. Bạn có thể sử dụng giá của nhà bán lẻ phổ biến làm hướng dẫn cho cửa hàng trực tuyến của riêng bạn. Công cụ mua sắm so sánh này cho phép bạn so sánh giá trực tuyến bằng cách lấy dữ liệu sản phẩm từ hàng ngàn cửa hàng trên web. Bạn có thể sử dụng một loạt các bộ lọc để thay đổi kết quả tìm kiếm của mình và so sánh giá trực tuyến để có được giao dịch trực tuyến tốt nhất.



**Pronto.com searches all the top stores,
so that you don't have to.**

When you shop here, you shop everywhere.

 × Find It

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

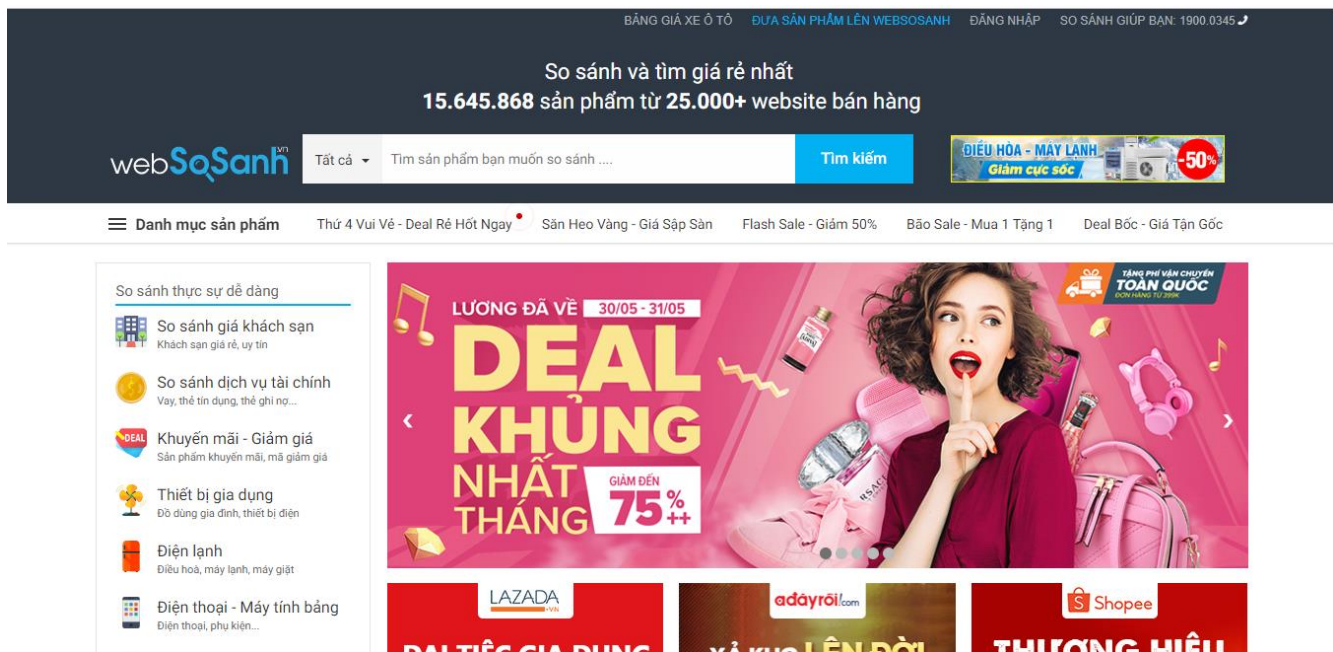
- Ở Việt Nam một vài năm trở lại đây, thương mại điện tử bắt đầu bùng nổ kéo theo sự xuất hiện của các dịch vụ so sánh giá. các trang web này thường sẽ thống kê dữ liệu từ các nhà bán hàng và đưa ra sắp xếp, gợi ý cho khách hàng mà không trực tiếp bán sản phẩm. Doanh thu sẽ đến từ các thỏa thuận chiết 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,...),...

2.5 So sánh một số website so sánh giá ở Việt Nam

- WebSoSanh.vn
 - Websosanh được xem là công cụ so sánh giá trực tuyến tốt nhất hiện nay. Với số lượng sản phẩm đa dạng và nhà cung cấp rộng khắp các tỉnh thành giúp bạn không chỉ so sánh với trang bán hàng trực tuyến mà còn có những địa chỉ ở gần nơi bạn ở. Hiện nay websosanh.vn có cơ sở dữ liệu gồm 6 triệu sản phẩm nên có thể nói là bao trùm tất cả các sản phẩm có mặt trên thị trường Việt Nam. Ngoài ra còn có 1 số tính năng so sánh giá cước taxi, giá các gói chụp ảnh cưới, giá chi phí thẻ tín dụng, so sánh giá bánh trung thu... thì chưa có có dịch vụ nào tương tự ở Việt Nam có được.
 - Mỗi khi nhập đầy đủ thông tin về một mặt hàng bạn đang quan tâm, các kết quả trả về của WebSoSanh.vn cũng được sắp xếp theo thứ tự giá từ thấp tới cao, các thông tin ở dưới cũng hiển thị cho bạn biết tình trạng còn hàng hay hết hàng và nếu quan tâm tới sản phẩm nào bạn có thể bấm vào "Tới nơi bán" để tìm hiểu thêm thông tin sản phẩm. Bạn cũng có thể sử dụng bộ lọc bên trái của trang để thu gọn lại kết quả tìm kiếm, các tùy chọn gồm vị trí địa lý, cửa hàng bán sản phẩm và lọc theo giá bán.
 - Ưu điểm:
 - Giao diện thân thiện: Ấn tượng đầu tiên khi bạn truy cập vào trang này là giao diện khá thân thiện và đẹp mắt. Mọi thứ đều rõ ràng và dễ thao tác tìm kiếm cũng như so sánh giá.
 - Nhiều thông tin hữu ích: Trang này có blog tin tức khá chuyên nghiệp và nhiều chuyên mục như đánh giá sản phẩm, tư vấn mua sắm, xu hướng thị trường và mẹo vặt. Đây là những thông tin tham khảo cần thiết để hiểu biết về sản phẩm trước khi mua sắm.

- Nhiều dịch vụ so sánh: Ngoài chức năng so sánh giá các sản phẩm mua sắm online thì trang này mở rộng sang so sánh giá khách sạn và các sản phẩm tín dụng. Tuy nhiên do không chuyên nên cũng có nhiều điểm hạn chế và khó sử dụng. Nhưng đây cũng là công cụ giúp bạn tham khảo rất tốt.
- Nhược điểm:
 - Giao diện quá nhiều thông tin: Ngoài tính thân thiện phía trên như menu và chuyên mục cũng như công cụ tìm kiếm rõ ràng. Bên dưới có quá nhiều thông tin khuyến mãi và link tiếp thị liên kết không cần thiết và gây khó chịu cho người cần muốn so sánh giá. Bạn nhìn kỹ thì giống như trang bán hàng chứ không phải đây là trang so sánh giá. Nhưng thực ra đây là những link tiếp thị sản phẩm dẫn tới nơi bán.
 - Nhiều nhà cung cấp: Quá nhiều đối tác giới thiệu sản phẩm trên trang này nếu không kiểm soát kỹ thì sẽ có rất nhiều sản phẩm kém chất lượng, thông tin quá nhiều gây rối cho người tìm kiếm.
 - Quảng cáo nhiều: Websosanh đang liên kết với rất nhiều đối tác bán lẻ cho nên nhưng đối tác nào chi tiền sẽ được nổi bật hơn và đề xuất phía trên. Vì vậy hãy xem xét cẩn thận trước khi mua.



- GoBear là công cụ chuyên so sánh 3 sản phẩm chính là bảo hiểm, thẻ tín dụng và vay tín chấp. Đúng là công cụ này rất hữu ích cho người có nhu cầu sử dụng các dịch vụ trên nhưng chưa rõ nên chọn cái nào là phù hợp hay tiết kiệm nhất. Gobear là đơn vị tiên phong trong việc so sánh các sản phẩm bảo hiểm và tín dụng cá nhân không chỉ ở Việt Nam mà còn ở các nước khác trong khu vực Đông Nam Á.

- Ưu điểm:
 - Giao diện đơn giản dễ sử dụng: Khi mới vào bạn sẽ thấy ngay những gì bạn cần và nó quá chi tiết và cực kỳ trực quan rất dễ sử dụng chỉ cần 1 2 cú click là bạn đã thao tác xong. Kết quả sẽ cho ra rất nhanh để bạn tham khảo và chọn đúng những gì bạn cần và phù hợp nhất.
 - So sánh tiện lợi: Sau khi tham khảo thông tin chi tiết từng ngân hàng thì bạn có thể so sánh bất kỳ từ 2 3 ngân hàng khác nhau để xem sự khác nhau và dễ dàng chọn lựa cái nào tối ưu hơn.
 - Thông tin hữu ích: Ngoài công cụ so sánh tiện lợi bạn còn có thể tham khảo rất nhiều thông tin hữu ích về bảo hiểm, thẻ tín dụng và vay tín chấp.
 - Blog chia sẻ kinh nghiệm: Có nhiều bài viết chia sẻ kiến thức và kinh nghiệm hay cho bạn tham khảo về các sản phẩm này.
- Nhược điểm:
 - Đối tác chưa nhiều: Các ngân hàng, công ty tài chính bảo hiểm đang hoạt động tại Việt Nam chưa có đầy đủ trên trang này. Cho nên bạn sẽ gặp khó khăn nếu ngân hàng, công ty bạn muốn so sánh lại không có thì gặp khó khăn cho bạn. Tuy nhiên mình thấy các ngân hàng lớn cũng tương đối đầy đủ, công ty bảo hiểm về du lịch thì cũng khá tốt, chỉ có vay tín chấp thì không có công ty tài chính mà chỉ có một số ngân hàng tham gia. Hay là tạo thẻ tín dụng cũng chỉ có một số ngân hàng chứ không nhiều.
 - Sản phẩm chưa đầy đủ: Các sản phẩm bảo hiểm, vay tín chấp, thẻ tín dụng chưa đầy đủ như mong đợi của nhiều người.
 - Bảo hiểm: Chỉ có dịch vụ bảo hiểm du lịch, trong khi bảo hiểm nhân thọ thì nhiều người quan tâm lại không có.
 - Thẻ tín dụng: Chỉ có một số ngân hàng, không có đầy đủ các ngân hàng để so sánh chính xác.
 - Vay tín chấp: Chỉ so sánh một số ngân hàng chứ không có các công ty tài chính tham gia

BH du lịch

Thẻ tín dụng

Vay tín chấp

Tôi đang tìm gói bảo hiểm [Theo chuyến ▾](#)

dành cho [cá nhân ▾](#) .

Tôi sắp đi đến [Đài Loan ▾](#) .

➕ Thêm một điểm đến

📅 [31-05-2019](#) đến ngày [09-06-2019](#)


XEM KẾT QUẢ

BẮT ĐẦU LẠI

HƠN 40 TRIỆU NGƯỜI DÙNG ĐÃ TIN TƯỞNG GOBEAR

Tìm kiếm. So sánh. Chọn lựa! Cách so sánh công bằng nhất.

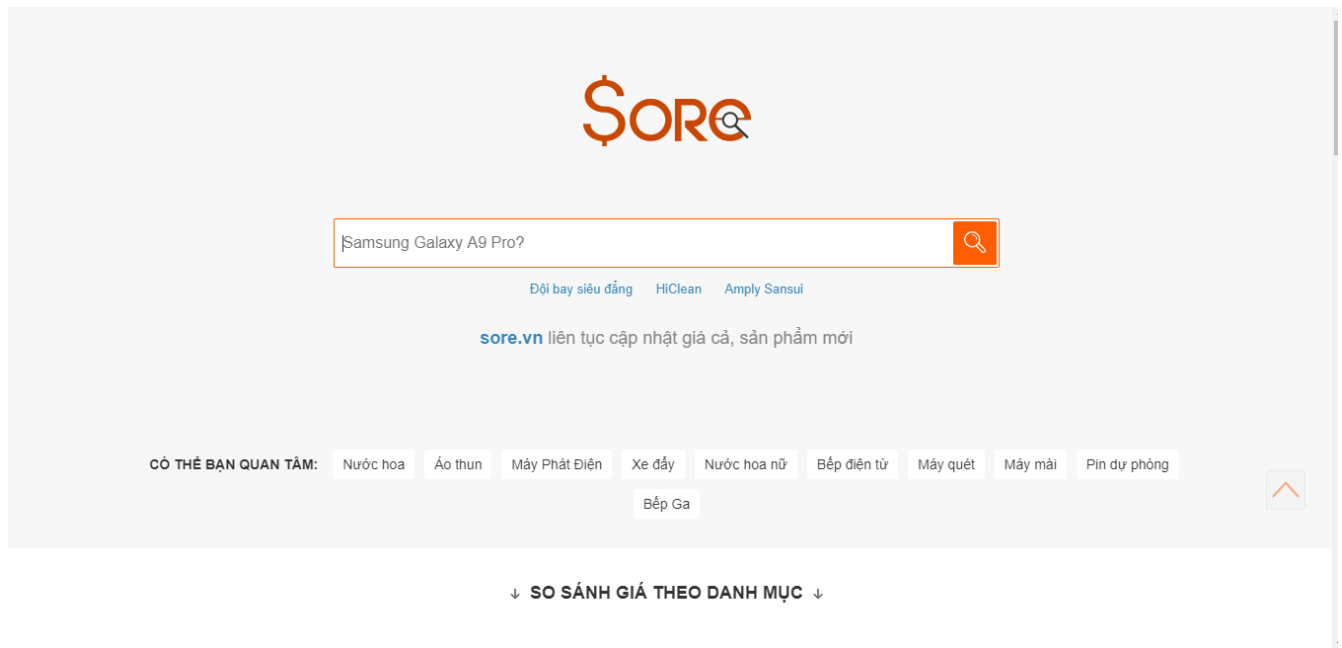
Tìm kiếm hàng trăm sản phẩm tài chính cá nhân từ những công ty uy tín



BIDV HOME citibank OceanA Bank

- Sore.vn

- Sore.vn website miễn phí giúp bạn tìm sản phẩm, so sánh giá rẻ các sản phẩm, mặt hàng và cửa hàng với nhau trên khắp cả nước. Mọi thông tin như giá cả các sản phẩm điện máy, điện thoại, đồ điện tử, điện lạnh, kỹ thuật số, thiết bị văn phòng, viễn thông, gia dụng, mẹ và bé, thời trang,... được thống kê giá một cách chính xác nhất từ các nhà bán lẻ lớn, nhỏ tại Việt Nam.
- Ưu điểm:
 - Nhiều đơn vị bán hàng: Có cả nhà bán lẻ trực tuyến lớn trên cả nước và nhà bán lẻ nhỏ tại địa phương nhiều tỉnh thành.
 - Giao diện menu tốt: Giao diện cũng trực quan dễ tìm kiếm danh mục sản phẩm.
- Nhược điểm:
 - Bảng so sánh giá chưa rõ ràng: Việc sắp xếp giá chưa hợp lý và chính theo tiêu chí giá. Sản phẩm đầu tiên có ghi chú quảng cáo thì đã hiểu còn sản phẩm thứ 2 giá cao hơn sản phẩm thứ 4 nhưng lại nằm trên.
 - Sản phẩm chưa xác nhận: Có nhiều sản phẩm chưa xác nhận không biết là chưa xác nhận đơn vị bán hàng hay xác nhận cái gì ở đây. Nhưng mà nếu chưa xác nhận rõ ràng thì không nên đăng lên hoặc hướng dẫn rõ người dùng cần chú ý hãy nếu mua thì như thế nào có ảnh hưởng gì không.



- Ngoài ra còn nhiều trang web so sánh giá khác như : sosanhgia.com , topgia.vn , atadi.vn , trivago.vn , Skyscanner.com.vn, Tripadvisor.com, Traveloka.com

2.6 Lợi ích của web so sánh giá

- Đối với người tiêu dùng
 - Website so sánh giá cho phép người tiêu dùng tìm kiếm sản phẩm mà mình đang quan tâm (qua thanh tìm kiếm hoặc danh mục sản phẩm) với kết quả trả ra là các gian hàng trực tuyến đang bán sản phẩm đó cùng mức giá bán tương ứng. Người tiêu dùng có thể dễ dàng lựa chọn nhà cung cấp có mức giá hợp lý nhất.
 - Danh mục sản phẩm trên các website so sánh giá phong phú hơn nhiều so với một đơn vị bán hàng trực tuyến thông thường. Người dùng gần như có thể tìm thông tin về bất kỳ sản phẩm nào khi sử dụng công cụ này.
 - Bên cạnh thông tin về giá, các website so sánh giá còn cung cấp cả thông tin về đặc tính của sản phẩm cũng như các bài viết đánh giá, hướng dẫn mua hàng với từng dòng sản phẩm, giúp người dùng đưa ra quyết định phù hợp hơn.
- Đối với các gian hàng trực tuyến
 - Website so sánh giá giúp gia tăng độ hiện diện trên Internet của các gian hàng trực tuyến, giúp nâng cao khả năng tiếp cận với khách hàng tiềm năng của các gian hàng này.
 - Biết được sản phẩm của mình có phổ biến không.
 - Biết được đối thủ cạnh tranh của mình và giá của họ.

- Xác định cách định giá cho sản phẩm

3 Nội dung đồ án:

3.1. Giới thiệu:

a) Lý do chọn đề tài:

- 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://vienthong.vn/> . Từ đó so sánh giá và đưa ra những gợi ý mua hàng cho người dùng.

b) 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, vienthong.com

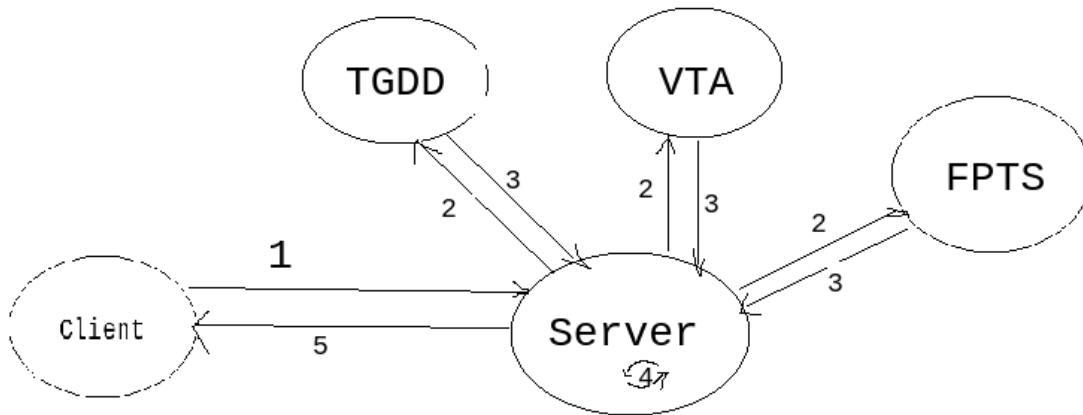
c) Tầm nhìn

- 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.

3.2 Hiện thực sản phẩm:

a) Mô hình xây dựng

- workflow của website:

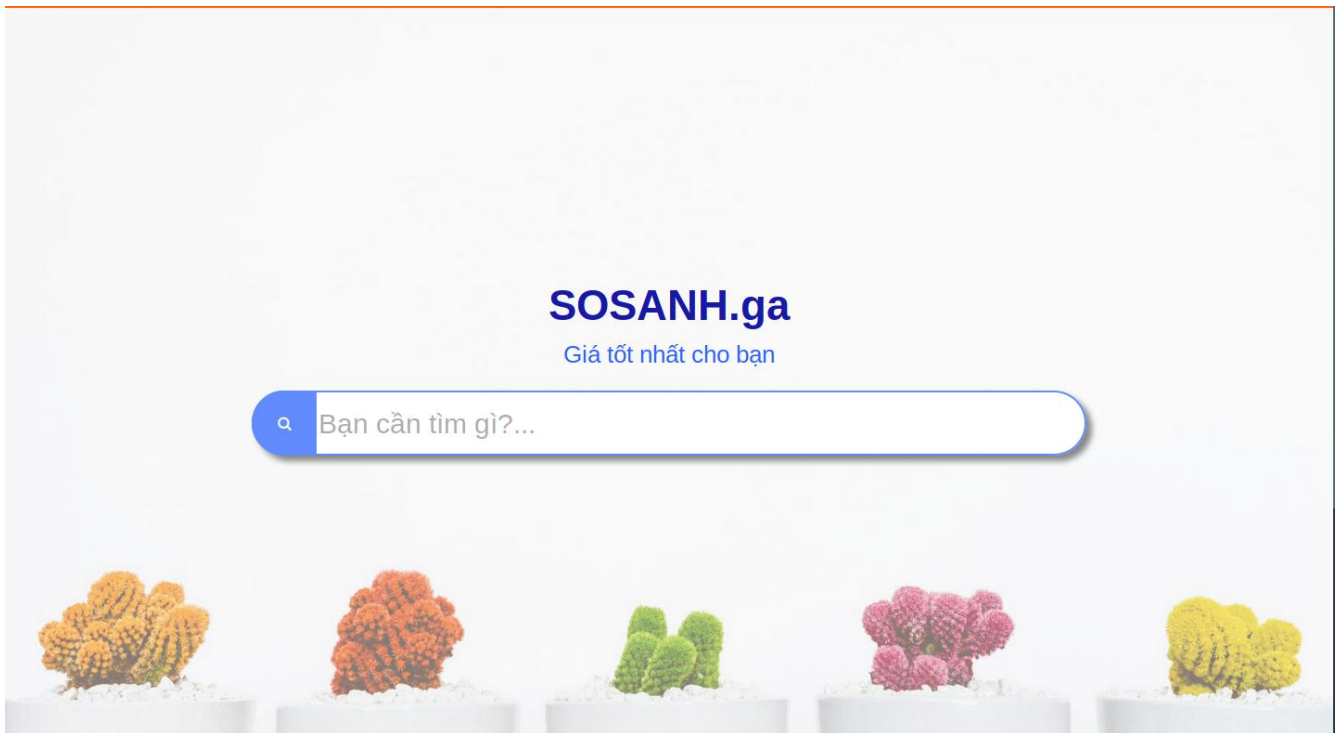


- Người dùng gửi yêu cầu là tên sản phẩm nhập từ ô tìm kiếm rồi gửi lên cho server xử lý.
- Server sẽ tiến hành đi crawl dữ liệu từ các trang nguồn với keyword mà client yêu cầu.
- Từ các dữ liệu nhận được từ các trang nguồn, server sẽ tiến hành kiểm tra các sản phẩm rồi sắp xếp theo thứ tự giá thấp đến cao.
- Server sẽ trả lại client response 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.

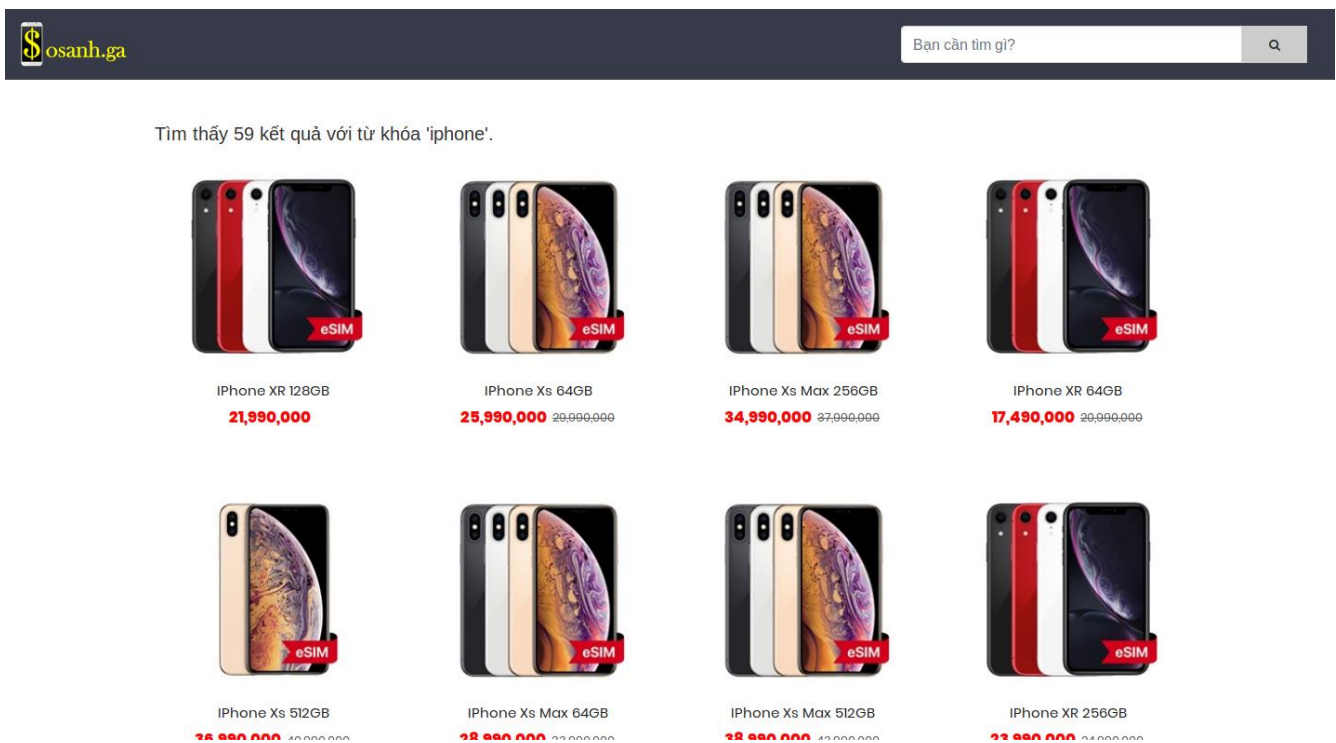
b) Xây dựng website

- 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.
- Xây dựng giao diện websites gồm 3 trang:

- Trang chủ:



- Trang hiển thị kết quả tìm kiếm:



- Trang chi tiết sản phẩm:

iPhone Xs 64GB



Màn hình : 5.8 inches, 1125 x 2436 Pixels
 Camera trước : 7.0 MP
 Camera sau : Dual Camera 12.0 MP
 RAM : 4 GB
 Bộ nhớ trong : 64 GB
 CPU : Apple A12 Bionic, 6, Đang cập nhật
 GPU : Apple GPU 4 nhân
 Dung lượng pin : Lâu hơn iPhone X 30'
 Hệ điều hành : iOS 12
 Thẻ SIM : eSIM và NanoSIM, 1 Sim

Giá tốt nhất tại: [FPT Shop](#)

Đến nơi bán

SO SÁNH GIÁ



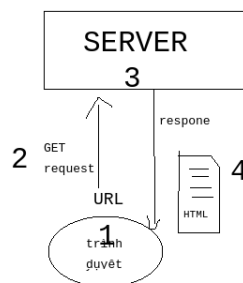
Đánh giá: ★★★★★
 25,990,000 29,990,000



Đến nơi bán

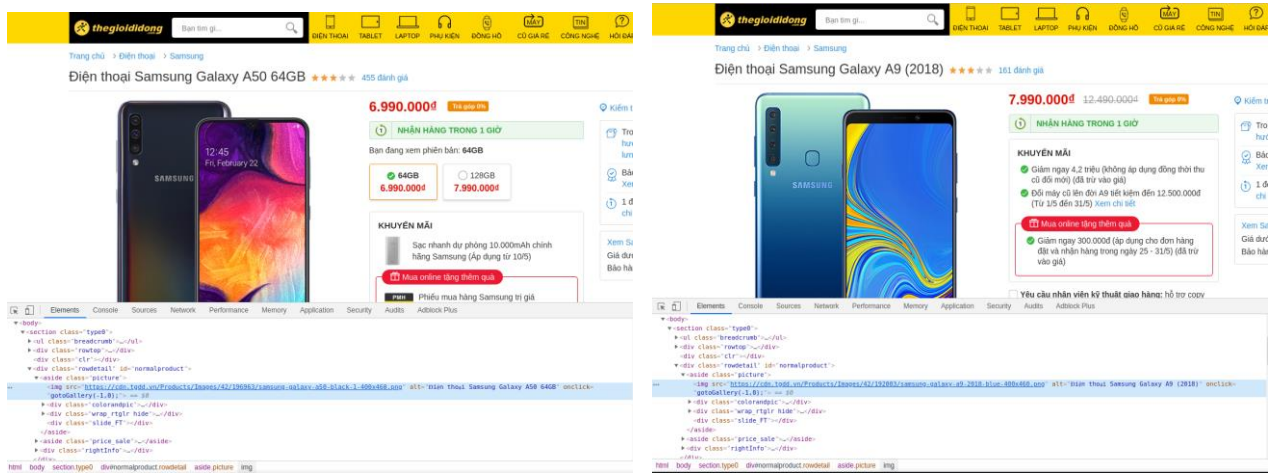
c) Xây dựng module crawl dữ liệu từ các trang bán hàng

- 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 là HTML.

- 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:
 - Client và server giao tiếp với nhau qua các http requests, clients (bình thường sẽ là trình duyệt) sẽ gửi request lên server với tên miền tương ứng (ví dụ thegioioidong.com), server sau khi xử lý yêu cầu được nhận sẽ trả về response là nội dung.
 - Với crawler, cần có công cụ để gửi request, nhận về response liên tục. Với python, có thư viện requests. Sau khi nhận về response, cần có thư viện bs4 để xử lý dữ liệu định dạng html.
 - Đầu tiên, khi người dùng nhập từ khóa và nhấn vào nút tìm kiếm, tiến hành gửi 3 get request đến 3 websites cần crawl với link tìm kiếm phù hợp và nhận về được 1 file dữ liệu html kết quả tìm kiếm cho mỗi trang. Bằng việc xem kết quả trả về, ta trích xuất được danh sách các sản phẩm tìm thấy rồi hiển thị ở trang danh sách kết quả.

ví dụ: link tìm kiếm của fptshop: <https://fptshop.com.vn/tim-kiem/{keyword}>

- Người dùng chọn vào một sản phẩm, sẽ có 1 get request tiếp theo lấy dữ liệu chi tiết sản phẩm của mỗi trang nguồn rồi tiếp tục xử và ta thu được các nội dung chi tiết của sản phẩm được chọn.

4.Sản phẩm - đánh giá

- Link websites: sosanh.ga
- Link source: <https://github.com/nguyenmao2101/sosanh.ga>
- Kết quả:

- Tìm hiểu được tầm quan trọng của web crawler, các kiến thức cơ bản và tự tạo ra được một sản phẩm ứng dụng công nghệ này.
- Có cái nhìn tổng quan về thị trường công cụ so sánh giá khu vực và thế giới, rút ra được hạn chế và chiến lược kinh doanh với thị trường tiềm năng này.
- Trang web so sánh giá giao diện đơn giản, thân thiện cung cấp tính năng cụ thể cũng là một bài luyện tập cho kỹ năng lập trình, phát triển ứng dụng web.
- Ưu điểm:
 - Giao diện đơn giản, tốc độ nhanh, tối ưu trên các thiết bị di động.
 - Giá sản phẩm được cập nhật chính xác theo thời gian thực.
- Hạn chế:
 - Với mức độ demo cho môn học nên websites hiện tại chỉ so sánh giá của những mặt hàng điện tử ở 3 cửa hàng là FPT shop, Thế giới di động và Viễn thông A.
 - Cấu trúc websites của các trang nguồn rất khác nhau kèm theo từng loại mặt hàng cũng có những cách bố trí khác nhau nên gây ra nhiều khó khăn trong quá trình xử lý và bảo trì nếu trang nguồn có thay đổi giao diện.
 - Chưa có các tính năng tối ưu trải nghiệm người dùng như tùy chọn so sánh, lọc sản phẩm, lưu sản phẩm,...

5. Hướng phát triển

- Với các công cụ so sánh giá với web crawler rất khó để có thể mở rộng quy mô mặt hàng, ngành hàng hoặc nhà phân phối. Cần lựa chọn sử dụng các biện pháp thay thế đơn giản hơn như: liên kết trang phân phối cung cấp dữ liệu sản phẩm định kỳ, yêu cầu được sử dụng hoặc xây dựng api để lấy dữ liệu theo thời gian thực,...
- Đối với các công cụ so sánh, như đã đánh giá ở trên, thị trường thương mại điện tử Việt Nam đang ngày càng lớn mạnh nên có rất nhiều hướng để phát triển:
 - Xây dựng công cụ với quy mô ngành hàng, mặt hàng rộng lớn cả sản phẩm lẫn dịch vụ, từ cả y tế đến giáo dục. Mở rộng số lượng nhà phân phối liên kết so sánh giá.
 - 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.
- Tiếp cận thị trường so sánh giá vào thời điểm này sẽ cũng không hẳn là một sự khởi đầu quá khó khăn bởi thị trường Việt Nam chỉ mới có vài tên tuổi nhỏ và rất nhiều loại ngành hàng chưa có công cụ so sánh phù hợp. Những điều cần có là một chiến lược tiếp cận, lựa chọn ngành hàng phù hợp, đối tượng khách hàng và quan trọng nhất mà rất nhiều sản phẩm chưa có được là trách nhiệm của công cụ so sánh giá đối với người dùng.

6. 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/>
- [11]<https://www.oberlo.com/blog/25-best-price-comparison-websites>