Project Proposal

Tên đề tài: Xây dựng Search Engine trên sàn thương mại điện tử Amazon

Thành viên nhóm:

- Nguyễn Minh Luân - 19110395
- Tô Thanh Phong - 19110050
- Tôn Thiên Thạch - 19110455
- Đào Quyết Phong – 19110427

1. Giới thiệu

Nhận thấy, ngành thương mại điện tử đang phát triển trong thời đại hiện nay. Một lượng dữ liệu khổng lồ được sinh ra từ các sản phẩm, mặt hàng rất đa dạng trên các nền tảng này. Điều đó tạo nên những mặt tốt để khai thác nguồn dữ liệu này cho đề tài của nhóm. Một trong những nơi có thể khai thác được nguồn dữ liệu từ các hoạt động trên đến từ các sàn giao dịch thương mại điện tử, chợ thương mại điển tử như Shopee, Lazada, Tiki,... nhưng website này có một lượng lớn người cung cấp sản phẩm (người bán) và người tiêu dùng (người mua) truy cập và giao dịch mỗi ngày. Từ đó, nhóm tôi quyết định chọn 01 trong những sàn thương mại điện tử tốt nhất hiện nay – sàn Amazon, làm đối tượng chính để khai thác dữ liệu.

Đề tài của nhóm là áp dụng kiến thức đã học, thực hành về information retrieval để xây dựng một Search Engine dùng Elasticsearch trên tập dữ liệu sản phẩm thu thập được trên Sàn thương mại điện tử Amazon (website – amazon.com) với kỹ thuật web scraping.

2. Dữ liệu

Giới thiệu về dữ liệu: hiện tại Amazon đang lưu trữ dữ liệu của **385 triệu sản phẩm** tiêu dùng (theo thống kê từ sellerengine.com). Với lượng sản phẩm rất lớn này Amazon đã chia nhỏ thành các mặt hàng (categories) để dễ dàng phân loại và tìm kiếm, với 30 mặt hàng sau:

Amazon Devices and Accessories	CDs and Vinyl	Health and Personal Care	Prime Video
Amazon Launchpad	Clothing	Home and Kitchen	Shoes and Bags
Apps and Games	Computers and Accessories	Jewellery	Software
Audible Books and Originals	Digital Music	Kindle Store	Sports and Outdoors
Automotive	DIY and Tools	Large Appliances	Stationery and Office Supplies
Baby Products	DVD and Blu-ray	Lighting	Toys and Games
Beauty	Electronics and Photo	Luggage	Watches
Books	Garden and Outdoors	Musical Instruments and DJ	
Business, Industry and Science	Gift Cards	PC and Video Games	
	Grocery	Pet Supplies	
	Handmade Products		

Danh mục 10 mặt hàng chiếm số lượng sản phẩm tiêu dùng cao nhất

(theo số liệu năm 2022 từ https://nuoptima.com/blog/amazon-product-categories)

- Home and Kitchen – 40%
- Sports and Outdoors – 21%
- Toys and Games – 19%
- Beauty and Personal Care – 19%
- Health, Household, and Baby Care – 18%
- Kitchen and Dining – 16%
- Office Products – 15%
- Garden and Outdoor – 14%
- Tools and Home Improvement – 14%
- Pet Supplies – 13%

Như vậy, giả sử chúng ta chỉ cần khai thác một mặt hàng "**Home & Kitchen**" (các sản phẩm gia dụng, bếp núc) thì cũng chiếm 40%, ***tức 154 triệu sản phẩm.***

Vì dữ liệu thu thập là các mặt hàng, sản phẩm nên đây sẽ là một số thuộc tính cơ bản mà nhóm sẽ thu thập từ Amazon:

TT	Nội dung	Mô tả
1	ASIN	Mã sản phẩm riêng biệt từ Amazon dùng để phân loại sản phẩm
2	Image	Hình ảnh về sản phẩm/Link hình sản phẩm
3	Title	Tiêu đề sản phẩm
4	Description	Mô tả về sản phẩm
5	Brand	Thương hiệu của sản phẩm
6	Star	Sao đánh giá từ user
7	Review Count	Đánh giá chất lượng từ người dùng
8	Price	Giá cả
9	Categories	Loại sản phẩm
10	URL	Link sản phẩm trên website

Dữ liệu thu thập về sẽ được lưu dưới dạng json, có thể theo mẫu nhóm hình dung như sau:


```
{
   "title": "Apple AirPods Pro (2nd Generation) Wireless Earbuds, Up to 2X More Active Noise Cancelling, Adaptive Transparency, Personalized Spatial Audio, MagSafe Charging Case, Bluetooth Headphones for iPhone",
   "asin": "B0BDHWDR12",
   "brand": "Apple",
   "stars": 4.7,
   "reviewsCount": 23653,
   "thumbnailImage":  "https://m.media-amazon.com/images/I/61f1YfTkTDL.__AC_SY445_SX342_QL70_FMwebp_.jpg",
   "breadCrumbs": "",
   "description": null,
   "price": {
     "value": 229,
     "currency": "$"
   },
   "url": "https://www.amazon.com/dp/B0BDHWDR12"
 },
```

3. Kế hoạch thực hiện:

Nhóm có kế hoạch thực hiện, phân công nhiệm vụ và đánh giá tiến độ các thành viên theo tuần như sau:

Tuần	Nhiệm vụ	Đánh giá theo tuần
01 (02/5 – 09/5/2023)	- Thu thập dữ liệu từ web bằng các thư viện, thuật toán. (Luân, Thạch) - Chuẩn bị thư mục data, thư mục image (Q. Phong) - Đặc tả ứng dụng, tính năng của Web Search (T. Phong)	
02 (09/5 – 16/5/2023)	- Chọn từ khóa, tạo chỉ mục (Q. Phong, Thạch) - Thiết kế giao diện Web Search Engine (T. Phong, Luân) - Viết báo cáo Report (Luân)	
03 (16/5 – 23/5/2023)	- Thiết kế Web Search Engine (Phong, Luân) - Kiểm thử (Cả nhóm) - Viết báo cáo Report (Luân) - Làm presentation (Q. Phong, Thạch)	
04 (23/5 – 30/5/2023)	- Hoàn thiện file report, presentation. (Cả nhóm) - Báo cáo cuối kỳ. (Cả nhóm)	

Ghi chú: *Nhóm có thể sẽ cập nhật thêm thông tin của Proposal cho đầy đủ, phù hợp hơn với bài Report cuối cùng trong quá trình tại các tuần. Cám ơn Thầy đã xem và lắng nghe.*