



Topic: Xây dựng Search Engine trên
sàn thương mại điện tử Amazon
Milestone – 16.05.2023

GVHD: Thầy Quách Đình Hoàng

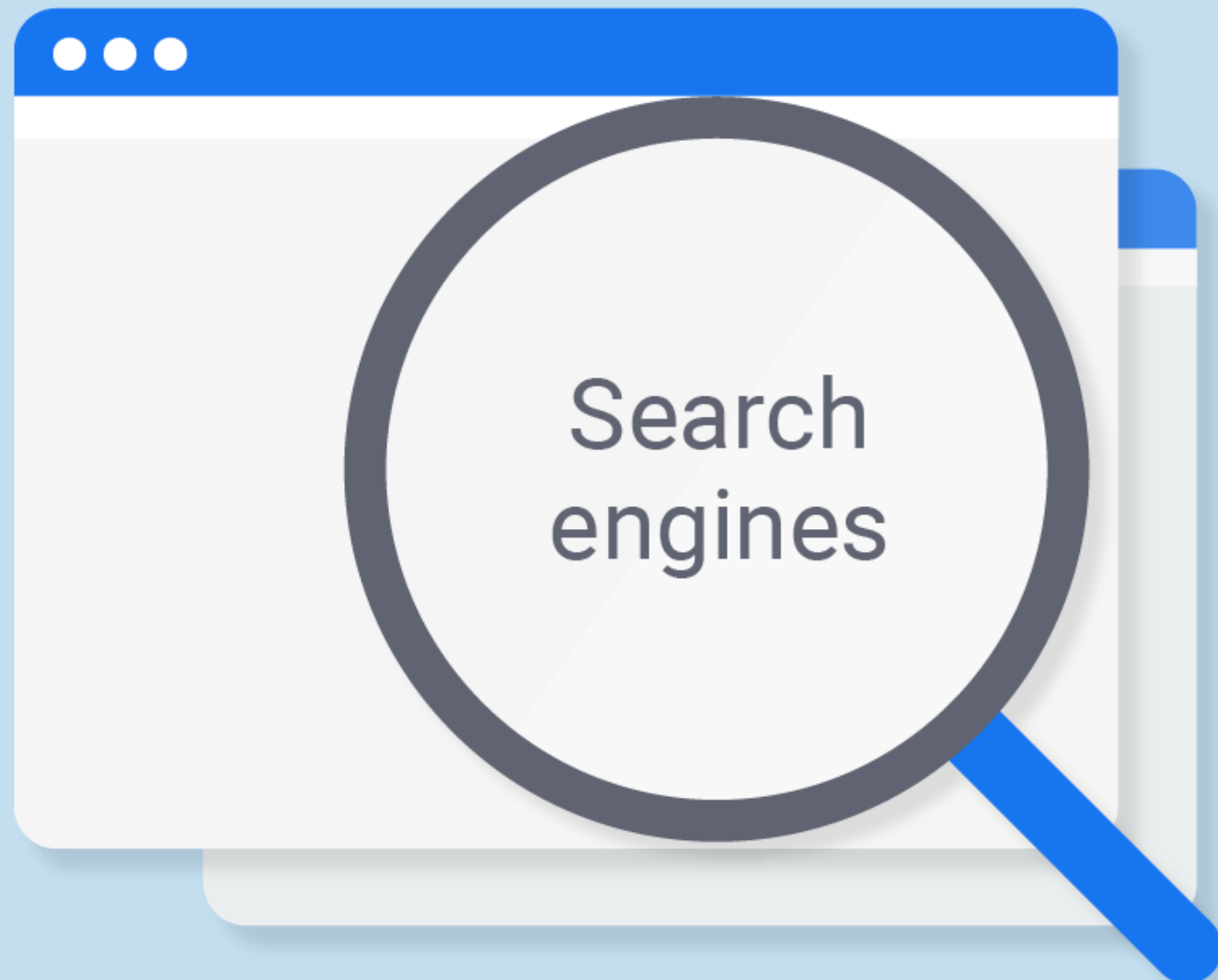
NHÓM 4 - SVTH:

Nguyễn Minh Luân - 19110395

Đào Quyết Phong – 19110427

Tô Thanh Phong - 19110050

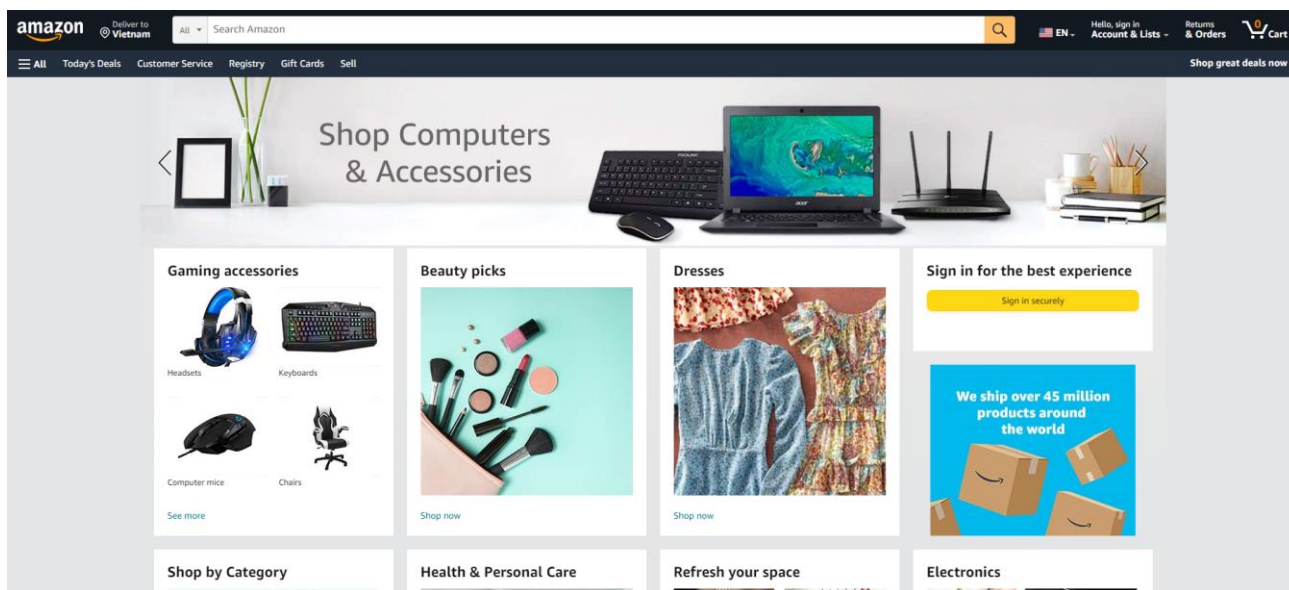
Tôn Thiên Thạch - 1911045



BÁO CÁO TIẾN ĐỘ

Tuần	Nhiệm vụ	Đánh giá theo tuần
01 (02/5 – 09/5/2023)	<ul style="list-style-type: none">- 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)- Viết Milestone	Hoàn thành
02 (09/5 – 16/5/2023)	<ul style="list-style-type: none">- Chọn từ khóa, tạo chỉ mục (Q. Phong, Thạch)- Thiết kế giao diện Web Search Engine (T. Phong, Luân)- Viết báo cáo Report (Luân)	Đang tiến hành
03 (16/5 – 23/5/2023)	<ul style="list-style-type: none">- Thiết kế Web Search Engine (Phong, Luân)- Kiểm thử (Cả nhóm)- Viết báo cáo Report (Luân)- Làm presentation (Q. Phong, Thạch)	Chưa hoàn thành
04 (23/5 – 30/5/2023)	<ul style="list-style-type: none">- Hoàn thiện file report, presentation. (Cả nhóm)- Báo cáo cuối kỳ. (Cả nhóm)	Chưa hoàn thành

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



Website Amazon.com

3 Key Ecommerce Trends You Should Know

FinancesOnline
REVIEWS FOR BUSINESS

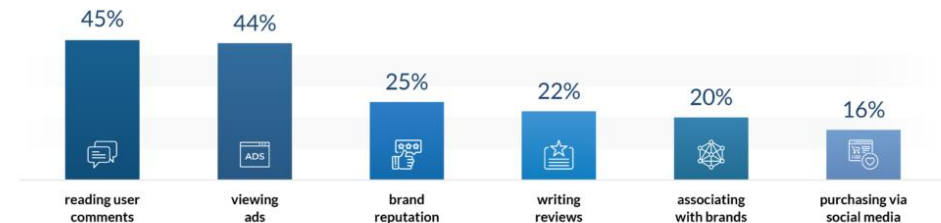
1 Global retail ecommerce sales will continue to soar

Global retail ecommerce sales figures, in trillion dollars:



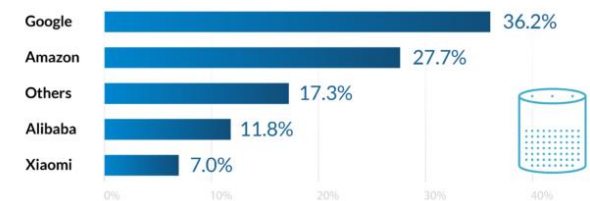
2 Social media activities increasingly influencing ecommerce retail behavior

Source: EcomStar



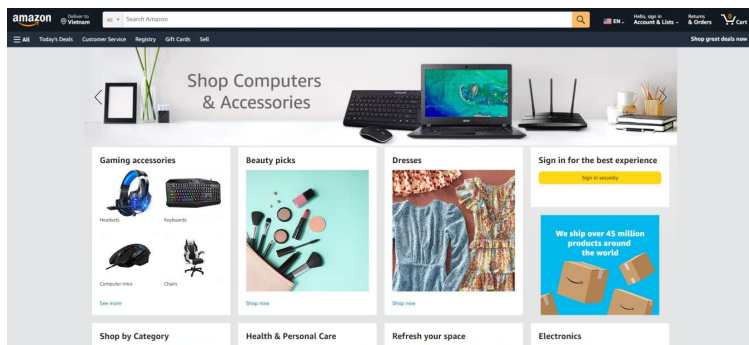
3 Smart speaker use for ecommerce among the big vendors will continue to rise

Source: Canalis



source: financesonline.com

TẬP DỮ LIỆU



(theo thống kê từ sellerengine.com)

Amazon đang lưu trữ
dữ liệu của **385**
triệu sản phẩm

- 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%

30 LOẠI MẶT HÀNG
- CATEGORY

Amazon Devices and Accessories
Amazon Launchpad
Apps and Games
Audible Books and Originals
Automotive
Baby Products
Beauty
Books
Business, Industry and Science

CDs and Vinyl
Clothing
Computers and Accessories
Digital Music
DIY and Tools
DVD and Blu-ray
Electronics and Photo
Garden and Outdoors
Gift Cards
Grocery
Handmade Products

Health and Personal Care
Home and Kitchen
Jewellery
Kindle Store
Large Appliances
Lighting
Luggage
Musical Instruments and DJ
PC and Video Games
Pet Supplies

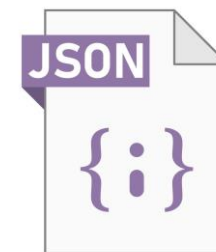
Prime Video
Shoes and Bags
Software
Sports and Outdoors
Stationery and Office Supplies
Toys and Games
Watches

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**

Các thuộc tính cơ bản cần thu thập:

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

```
{  
  "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"  
},
```



HIỆN THỰC VÀ KẾT QUẢ SƠ BỘ

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.select import Select
from webdriver_manager.chrome import ChromeDriverManager
import time
import random

URL = "https://www.amazon.com/"
web = webdriver.Chrome(executable_path=ChromeDriverManager().install())
web.get(url=URL)
```

#get categories

HIỆN THỰC VÀ KẾT QUẢ SƠ BỘ

```
1 #get categories
2 _search_dropdown = Select(web.find_element(By.ID, 'searchDropDownBox'))
3 categories = _search_dropdown.options
4 # convert WebElement to String
5 categories = [category.get_attribute("innerHTML") for category in categories]
6 print(categories)
```

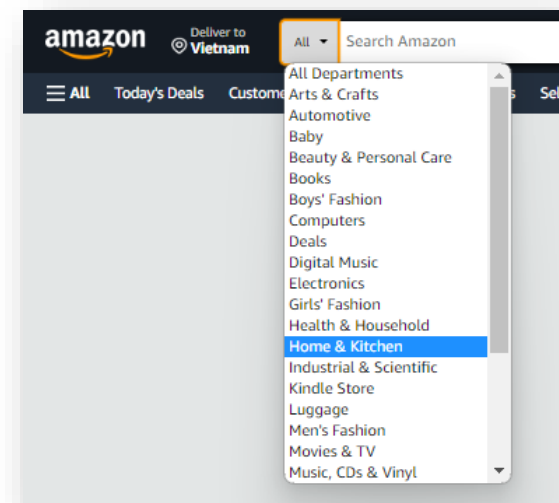
✓ 0.6s

Python

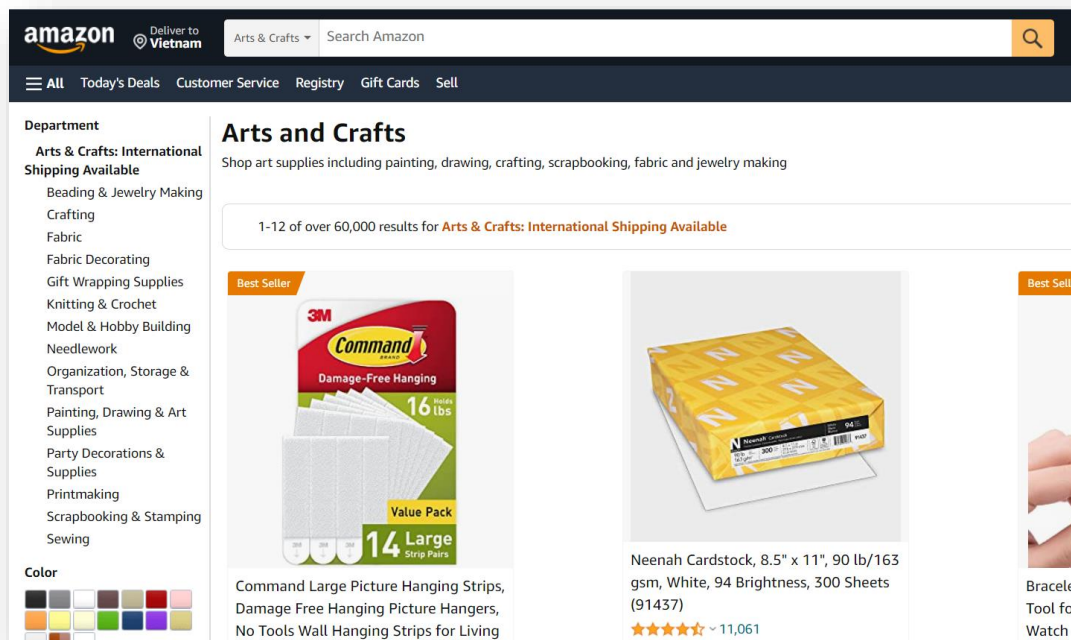
```
['All Departments', 'Arts & Crafts', 'Automotive', 'Baby', 'Beauty & Personal Care', 'Books', "Boys' Fashion", 'Computers', 'Deals', 'Digital Music', 'Electronics', "Girls' Fashion", 'Health & Household', 'Home & Kitchen', 'Industrial & Scientific', 'Kindle Store', 'Luggage', "Men's Fashion", 'Movies & TV', 'Music, CDs & Vinyl', 'Pet Supplies', 'Prime Video', 'Software', 'Sports & Outdoors', 'Tools & Home Improvement', 'Toys & Games', 'Video Games', "Women's Fashion"]
```

#get categories

#redirect to respective product page

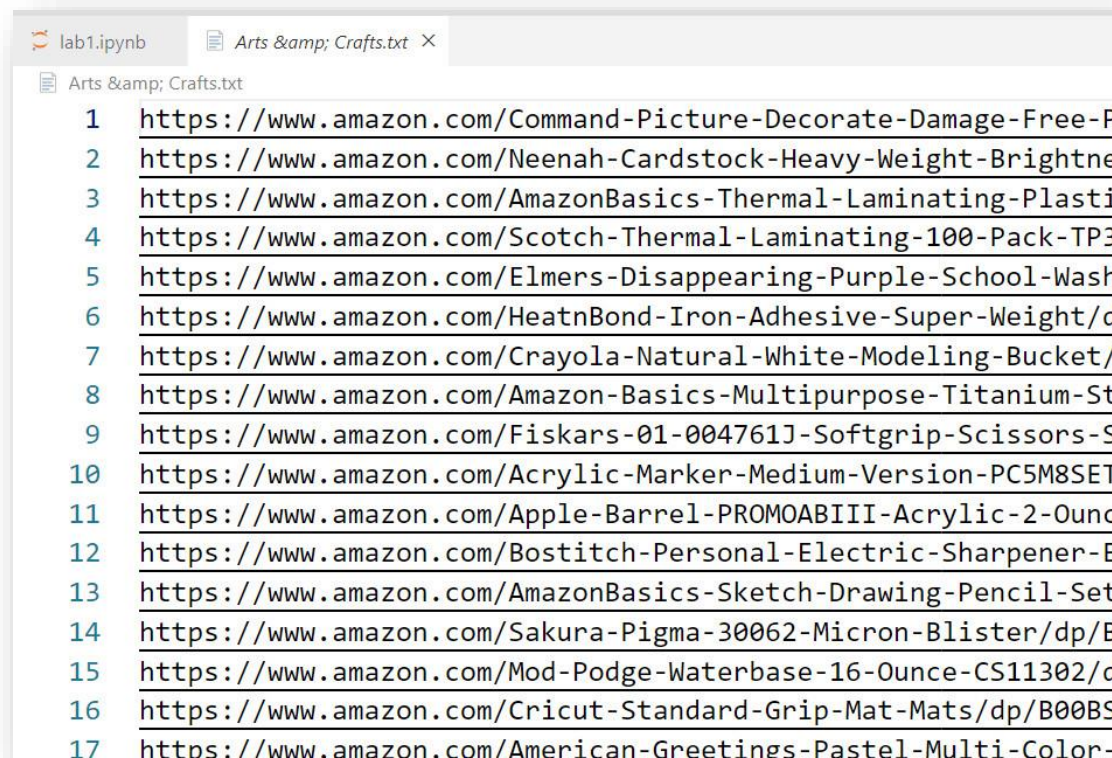


HIỆN THỰC VÀ KẾT QUẢ SƠ BỘ



#redirect to respective product page

#get all link of products



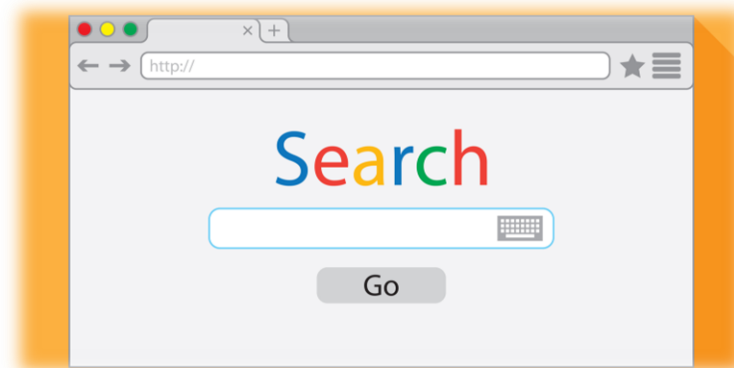
ĐẶC TẢ Web Search Engine

Chức năng:

- Tìm kiếm dữ liệu
- Tìm kiếm dữ liệu tùy chỉnh
- Lọc dữ liệu bằng bộ lọc
- Dowload dữ liệu đã cào
- Gợi ý các từ khóa liên quan khi tìm kiếm
- Gợi ý sản phẩm liên quan
- Tính năng xếp hạng và đánh giá kết quả tìm kiếm
- Tìm kiếm dựa trên thông tin từ trang web khác
- Phân tích dữ liệu
- Lưu trữ các tìm kiếm trước đây

Thành phần:

- Thanh tìm kiếm dữ liệu
- Bộ lọc dữ liệu
- Nút dowload dữ liệu
- Mục gợi ý sản phẩm
- Mục xếp hạng và đánh giá kết quả tìm kiếm
- Nút thêm thông tin từ trang web hoặc từ máy tính
- Mục hiển thị thông tin dữ liệu được phân tích
- Mục lưu trữ lịch sử tìm kiếm
- Thanh Nav hiển thị các chức năng của trang web



Không crawl hết data các Page của Web



Aw, Snap!

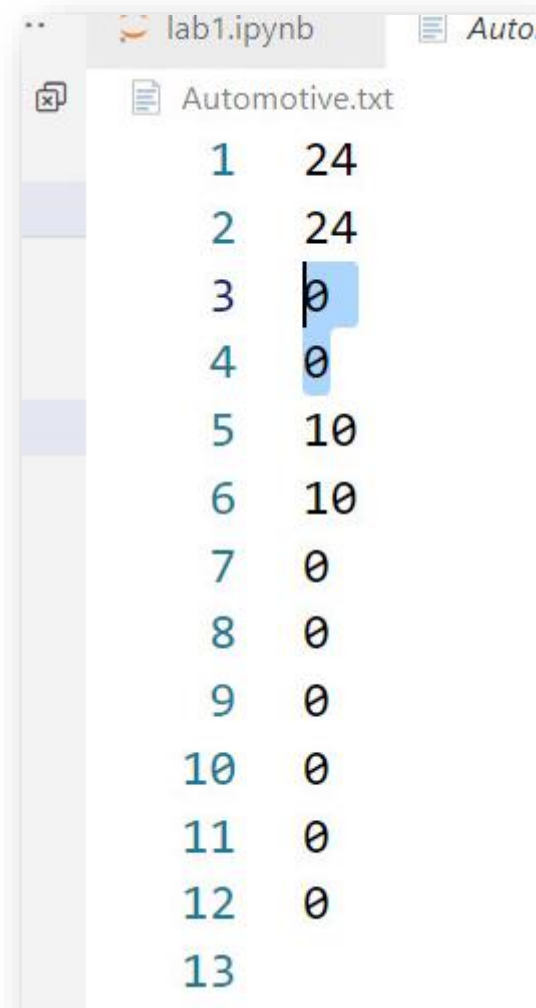
Something went wrong while displaying this webpage.

Error code: Out of Memory

[Learn more](#)

Thu thập được 70/400 Page

**Không lấy hết được
link dữ liệu**



Automotive.txt	
1	24
2	24
3	0
4	0
5	10
6	10
7	0
8	0
9	0
10	0
11	0
12	0
13	

KẾ HOẠCH TIẾP THEO

02 (16/5 – 23/5/2023)	<ul style="list-style-type: none">- Chọn từ khóa, tạo chỉ mục (Q. Phong, Thạch)- Thiết kế giao diện Web Search Engine (T. Phong, Luân)- Thực nghiệm với Elasticsearch- Viết báo cáo Report (Luân)	Đang tiến hành
--	--	-----------------------