

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

Khoa Hệ Thống Thông Tin

---oOo---



**BÁO CÁO LAB 3**

Giảng viên hướng dẫn : **ThS. Trình Trọng Tín**

Sinh viên : Nguyễn Đức Thành

MSSV : 23521449

Lớp: EC312.Q11

*Thành phố Hồ Chí Minh, ngày 03 tháng 11 năm 2025*

## Mục lục

1.	Yêu cầu đề bài: .....	1
2.	Giải thích mã nguồn .....	2
1.1	Mục tiêu & Tổng quan .....	2
	<b>Mục tiêu:</b> .....	2
	<b>Tổng quan:</b> .....	2
1.2	Thư viện & Vì sao dùng .....	2
1.3	Cấu hình & Hằng số .....	2
1.4	Lớp Product (OOP) .....	2
1.5	Session + Retry .....	2
1.6	Hàm tải & Parse JSON-LD .....	3
1.7	Phân tích trang chi tiết sản phẩm .....	3
1.8	Selenium & phân nhân 'Xem thêm' .....	3
1.9	Đa luồng & Giới hạn tốc độ .....	3
1.10	Hàm main() & luồng hoạt động .....	3
1.11	Thiết kế & Ưu điểm .....	3
1.12	Rủi ro & Cách xử lý .....	4
3.	Hình ảnh thu được .....	5



## 1. Yêu cầu đề bài:

1. Xây dựng Tiki Crawler để lấy thông tin sản phẩm theo danh mục cụ thể, ví dụ: Máy tính xách tay, Chăm sóc da, Điện thoại di động,... (5đ) (Sinh viên có thể chọn Hasaki, CellphoneS, ...)
2. Tạo "Mẫu nhập sản phẩm" cho WooCommerce. Tham khảo sample\_dataset.txt. (5đ)

### ***Yêu cầu nộp:***

- File report có hình chụp các mã nguồn, quá trình chạy crawler.
- Đính kèm mã nguồn của Crawler

## 2. Giải thích mã nguồn

### 1.1 Mục tiêu & Tổng quan

#### Mục tiêu:

- Mở trang danh mục Tiki bằng Selenium.
- Bấm nút “Xem thêm” nhiều lần để tải thêm hàng hóa.
- Thu thập URL sản phẩm từ trang.
- Dùng requests + BeautifulSoup để lấy chi tiết sản phẩm theo chuẩn WooCommerce.

#### Tổng quan:

- Selenium: tải nội dung động, click “Xem thêm”.
- Requests + BeautifulSoup: parse dữ liệu sản phẩm.
- Multi-thread + batching: thu thập nhanh nhưng tránh bị chặn.

### 1.2 Thư viện & Vì sao dùng

- selenium: xử lý nội dung động và thao tác nút “Xem thêm”.
- requests: tải trang HTML tĩnh nhanh gọn.
- BeautifulSoup: phân tích HTML, tìm JSON-LD dữ liệu sản phẩm.
- concurrent.futures: xử lý đa luồng.
- urllib.parse: chuẩn hóa URL.
- pandas: lưu dữ liệu ra file CSV.

### 1.3 Cấu hình & Hằng số

WOO\_HEADERS: danh sách các cột theo định dạng WooCommerce.

USER\_AGENTS: danh sách nhiều trình duyệt khác nhau, tránh bị phát hiện bot.

### 1.4 Lớp Product (OOP)

Mục tiêu: tạo một đối tượng sản phẩm thống nhất với các trường dữ liệu cố định.

Ưu điểm: dễ chuyển đổi thành dictionary và ghi ra CSV.

### 1.5 Session + Retry

Hàm build\_session() tạo session có headers, retry 3 lần khi gặp lỗi 429, 500, 503,...

Giúp tăng độ bền khi truy cập nhiều lần.

## 1.6 Hàm tải & Parse JSON-LD

`parse_ldjson_all()`: trích xuất dữ liệu từ script `application/ld+json`.

`pick_product_obj()`: chọn đối tượng có `@type = "Product"`.

`availability_to_bool()`: chuyển availability URL sang True/False.

## 1.7 Phân tích trang chi tiết sản phẩm

Hàm `build_product_from_url()`:

- Tải trang chi tiết.
- Lấy thông tin từ JSON-LD và meta tags.
- Map dữ liệu sang định dạng WooCommerce.

## 1.8 Selenium & phần nhấn 'Xem thêm'

`scrape_category_urls()`:

- Mở trình duyệt Chrome.
- Nhấn nút “Xem thêm” `times_click_more` lần.
- Thu thập các thẻ `<a class='product-item'>`.
- Làm sạch và loại trùng URL.

## 1.9 Đa luồng & Giới hạn tốc độ

`fetch_all_products()`: chạy 8 luồng song song.

`process_in_batches()`: chia batch 240 URL, sau mỗi batch nghỉ 31 giây.

## 1.10 Hàm main() & luồng hoạt động

`main()`:

1. Gọi `scrape_category_urls` để lấy danh sách URL.
2. Chạy `process_in_batches` để lấy dữ liệu chi tiết từng sản phẩm.
3. Lưu dữ liệu ra file `tiki_product_crawling.csv`.

## 1.11 Thiết kế & Ưu điểm

- Sử dụng JSON-LD → dữ liệu ổn định, tránh phụ thuộc DOM.
- Có batch giới hạn → tránh bị Tiki chặn.
- Đa luồng hợp lý → tăng tốc độ mà vẫn an toàn.

### 1.12 Rủi ro & Cách xử lý

- 429 Too Many Requests: đã có batch + cooldown.
- DOM thay đổi: cần cập nhật XPath hoặc dùng data-view-content.
- Dữ liệu thiếu: nếu không có JSON-LD thì fallback về meta description.

### 3. Hình ảnh thu được

#### Output của code

```
Clicked 'Xem thêm' (1/20)
Clicked 'Xem thêm' (2/20)
Clicked 'Xem thêm' (3/20)
Clicked 'Xem thêm' (4/20)
Clicked 'Xem thêm' (5/20)
Clicked 'Xem thêm' (6/20)
Clicked 'Xem thêm' (7/20)
Clicked 'Xem thêm' (8/20)
Clicked 'Xem thêm' (9/20)
Clicked 'Xem thêm' (10/20)
Clicked 'Xem thêm' (11/20)
Clicked 'Xem thêm' (12/20)
Clicked 'Xem thêm' (13/20)
Clicked 'Xem thêm' (14/20)
Clicked 'Xem thêm' (15/20)
Clicked 'Xem thêm' (16/20)
Clicked 'Xem thêm' (17/20)
Clicked 'Xem thêm' (18/20)
Clicked 'Xem thêm' (19/20)
Clicked 'Xem thêm' (20/20)
Collected 841 unique product URLs.

== Batch 1/4: 240 URLs ==
Hit 240 requests. Cooling down 31s to respect Tiki limit...

== Batch 2/4: 240 URLs ==
[ERR] Exceeded 30 redirects.
[ERR] Exceeded 30 redirects.
[ERR] Exceeded 30 redirects.
[ERR] Exceeded 30 redirects.
```

```
Clicked 'Xem thêm' (13/20)
Clicked 'Xem thêm' (14/20)
Clicked 'Xem thêm' (15/20)
Clicked 'Xem thêm' (16/20)
Clicked 'Xem thêm' (17/20)
Clicked 'Xem thêm' (18/20)
Clicked 'Xem thêm' (19/20)
Clicked 'Xem thêm' (20/20)
Collected 841 unique product URLs.

== Batch 1/4: 240 URLs ==
Hit 240 requests. Cooling down 31s to respect Tiki limit...

== Batch 2/4: 240 URLs ==
[ERR] Exceeded 30 redirects.
[ERR] Exceeded 30 redirects.
[ERR] Exceeded 30 redirects.
[ERR] Exceeded 30 redirects.
[ERR] Exceeded 30 redirects.
[ERR] Exceeded 30 redirects.
[ERR] Exceeded 30 redirects.
[ERR] Exceeded 30 redirects.
Hit 240 requests. Cooling down 31s to respect Tiki limit...

== Batch 3/4: 240 URLs ==
Hit 240 requests. Cooling down 31s to respect Tiki limit...

== Batch 4/4: 121 URLs ==

Saved 833 products to tiki_product_crawling.csv
```



Các dòng đầu file csv thu được:

ID	Type	SKU	Name
277874345	simple	7694863176791	Đèn Pha LED Ô Tô - Xe Máy H4 36W: Siêu Sáng Với 3 Nhiệt Màu AUTO Tiết Kiệm Năng Lượng Loại Tốt
276853295	simple	9082132007232	Hạt giống Xà Lách các loại, nảy mầm tốt, dễ trồng, nhanh thu hoạch, cây khỏe - Hạt giống Rạng Đông
	simple		
276950025	simple	2247601019617	Bộ hộp cơm giữ nhiệt LocknLock Tower Vacuum Lunch Box - LHC8058 - Hàng chính hãng
277860856	simple	9194721960038	Thiết Bị Tiết Kiệm Điện Thông Minh Nhật Bản – Bí Kíp Vàng Cho Mọi Gia Đình Việt
276666250	simple	1330109746205	Khăn giấy cuộn đa năng lau bếp bát đĩa đồ gia dụng kích thước 20x25cm thấm hút tốt, dai khó rách, có thể tái sử
276409127	simple	9774799581980	Giấy Đựng Sách Để Bàn Deli - Kệ Chia 4 Ngăn Văn Phòng Học Sinh Sinh Viên - Chặn Sách Nhiều Ngăn Giúp Sắp Xế
276463998	simple	3357710749518	COMBO 3 Bóng Giặt Silicon Cầu Gai Chống Nhấn Quần Áo Quả Xanh To 5.5CM Hamart Giặt Sạch Vượt Trội
276440259	simple	6025853640672	Bình Gas Bơm Bật Lửa Bluesky 5 Đầu Bơm Dung Tích 150gr/275ml Siêu Tiện Dụng, Bình Bơm Gas Bật Lửa Hàn Qu
277470440	simple	2690621485997	Bình Nước Thuỷ Tinh Kèm Nắp Pha Trà Thông Minh Cao Cấp Có Lọc Trà Inox Sang Trọng Tặng Túi Chống Nóng Qu
276025123	simple	7855665961646	Chảo inox 3 lớp nguyên khối Elmich Trimax EL 3846OL16 size 16cm
277835691	simple	6733441872832	Nồi phủ sứ chống dính Elmich Olive Classic EL- 5535OV - Hàng chính hãng
275929239	simple	4913952912515	Bình nước thủy tinh Elmich EL-8350T052 dung tích 520ml
277579187	simple	9235158882658	Nắp Đậy Cống Thoát Nước Chống Mùi Hối, Côn Trùng - Chuyên Dùng Cho Sàn Nhà Tắm, Nhà Vệ Sinh - HÀNG CHỈ
275246659	simple	2655185633747	Nệm Trải Sàn, Đệm, Topper Hàng Xuất Nhật 1Mx2M, 1M2X2M, 1M4x2M, 1M6x2M, 1M8X2M Dày 9CM Tặng Kèm Dr
275374645	simple	9157060773345	Cốc giữ nhiệt inox 304 Elmich EL8345 dung tích 480ML - Hàng Chính Hãng
275059035	simple	5983279524305	Bình giữ nhiệt Metro Cafe Tumbler LocknLock LHC4357 - Dung tích 500ml - Phủ sứ
275570185	simple	9902790305117	Bình đựng thức ăn giữ nhiệt inox 304 Elmich EL8336 dung tích 800ml
274823198	simple	8748684264070	Ổ Cắm Điện Mở Rộng Xoay 180 Độ
274957678	simple	8267328393679	Bình giữ nhiệt Flexi Sunhouse 920 ml Sunhouse KS-TU920
273247047	simple	1558233995914	Cốc giữ nhiệt inox 304 Elmich EL8309 dung tích 900ml - Hàng Chính Hãng
272484435	simple	2728990108038	Đồng hồ treo tường kim trôi không ồn minimalist phong cách tối giản trang trí phòng khách Bảo hành 12 tháng Del
271079046	simple	9170111379805	Đũa gỗ Nhật Bản vì 5 đôi KOKUSAI NDCK
253356473	simple	5739043260977	Pin sạc 9V Beston Li-Ion dung lượng cao, bộ sạc pin vuông dùng cho micro, đồng hồ vạn năng, chuông báo động,
249970011	simple	5133998734604	Chảo inox 3 lớp Elmich EL-3846OL Size 24cm
246167979	simple	1924384847270	Cây chà sàn nhà tắm, chổi cọ sàn và gạt nước 2 trong 1, cọ sạch khe hẹp và góc tường – Parroti Easy ES04
276121888	simple	1854346110106	Ống hút inox LocknLock loại ống hút thẳng, ống hút cong, cọ rửa ống hút - JoyMall
226036802	simple	4185734327530	Bình Giữ Nhiệt LocknLock 550ml LHC3249
275892618	simple	3924489285338	Khăn diều treo tường 1280 tờ/4 lớn Tondia duna rút tiền lối cao cấp tiết kiệm an toàn kèm móc treo (Thùng 4 gói)

column 5	column 6	column 7	column 8	
Published	Is featured?	Visibility in catalog	Short description	Description
1	0	visible		ĐẶC BIỆT : Không cần độ chế , cắt dây , chỉ cần cắm vào thay bóng như đèn zin của xe
1	0	visible		HẠT GIỐNG RẠNG ĐÔNG – GIỐNG TỐT CHO MÙA BỘI THU. HƯỚNG DẪN TRỒNG C
1	0	visible		
1	0	visible		Nắp vận với hệ thống Interlock: Dễ dàng lưu trữ đồ ăn, vận chuyển nhiều hộp cùng lúc
1	0	visible		Thiết Bị Tiết Kiệm Điện Nhật Bản Cao Cấp – Cách Giảm Điện Hiệu Quả 20% Mà Bạn Nế
1	0	visible		Khăn giấy cuộn đa năng lau bếp bát đĩa đồ gia dụng kích thước 20x25cm thấm hút tốt,
1	0	visible		TÍNH NĂNG SẢN PHẨM:Giúp lưu trữ sách vở tài liệu, tiết kiệm không gian bàn học bàn
1	0	visible		Bóng Giặt Silicon Cầu Gai Chống Nhấn Quần Áo Quả Xanh To 5.5CM Hamart Giặt Sạch
1	0	visible		- Đối với những loại bật lửa đất tiền bạn nên sử dụng những loại Gas tốt không tạp chấ
1	0	visible		Bình nước thủy tinh 750,1000,1200,1500ml tiện lợi, có lọc bã trà, tặng kèm túi vải cách
1	0	visible		Đặc điểm sản phẩmCông nghệ mới 3 lớp liên thân bền bỉ thách thức thời gian: Cấu tạo
1	0	visible		Câu Chuyện Sản Phẩm Olive – biểu tượng của sự trường tồn và sức khỏe, luôn gắn liền
1	0	visible		Đặc điểm nổi bậtThiết kế thời trang, hiện đại Chất liệu thủy tinh cao cấp, tuyệt đối an t
1	0	visible		Nắp Đậy Cống Thoát Nước Chống Mùi Hối, Côn Trùng - Chuyên Dùng Cho Sàn Nhà Tẩ
1	0	visible		CAM KẾT: Cam kết đổi trả nếu hàng không giống mô tả và hoàn tiền 100% nếu sản phẩ
1	0	visible		Chất Liệu Inox 304 – Bền Bỉ và An Toàn:Thân cốc được làm từ inox 304, chất liệu chón
1	0	visible		- Bình giữ nhiệt Metro Cafe Tumbler LHC4357 LocknLock - Dung tích 500ml - Thép ki
1	0	visible		Đặc điểm nổi bậtThiết kế nhỏ gọn, tinh tếChất liệu inox 304 tuyệt đối an toàn cho sức l
1	0	visible		Đóng gói: 1 ổ điện mở rộng (3 phích cắm) Tính năng Ổ điện mở rộng với 3 phích cắm ẹ
1	0	visible		Thiết kế trẻ trung, hiện đại với quai xách tiện lợiBình giữ nhiệt Flexi Sunhouse 920 ml S
1	0	visible		Thông tin sản phẩm– Thiết kế nắp bật chắc chắn, nước ko bị đổ ra ngoài, màu sắc cá t
1	0	visible		Đồng hồ treo tường kim trôi không ồn minimalist phong cách tối giản trang trí phòng ki
1	0	visible		Thương hiệu: KokusaiĐặc điểm sản phẩm:Đũa gỗ Nhật Bản vì 5 đôi Kokusai 23.5cm NĐ
1	0	visible		Pin sạc 9V Beston Li-Ion dung lượng cao, bộ sạc pin vuông dùng cho micro, đồng hồ v
1	0	visible		Chất liệu inox 304 cao cấp, tuyệt đối an toànLà dòng chảo không có sơn chống dính ni
1	0	visible		Chổi cọ sàn vệ sinh Parroti Easy ES04 được thiết kế thông minh, đầu cọ cực kỳ linh ho:
1	0	visible		Ống hút inox Lock&Lock loại ống hút thẳng, ống hút cong, cọ rửa ống hút - JoyMallKíc
1	0	visible		Đặc điểm nổi bậtBình Giữ Nhiệt LocknLock 550ml LHC3249 với thiết kế nắp mở một c
1	0	visible		- Nguyên liệu: 100% bột gỗ nguyên chất- Sản phẩm 4K: Không chất tẩy trắng (hàm lư

	column 10	column 11	column 12	column 13	column 14	column 15	column 16	
1	Date sale price starts	Date sale price ends	Tax status	Tax class	In stock?	Stock	Low stock amount	Bac
2		2025-12-03	taxable		1			no
3		2025-12-03	taxable		1			no
4			taxable					no
5		2025-12-03	taxable		1			no
6		2025-12-03	taxable		1			no
7		2025-12-03	taxable		1			no
8		2025-12-03	taxable		1			no
9		2025-12-03	taxable		1			no
10		2025-12-03	taxable		1			no
11		2025-12-03	taxable		1			no
12		2025-12-03	taxable		1			no
13		2025-12-03	taxable		1			no
14		2025-12-03	taxable		1			no
15		2025-12-03	taxable		1			no
16		2025-12-03	taxable		1			no
17		2025-12-03	taxable		1			no
18		2025-12-03	taxable		1			no
19		2025-12-03	taxable		1			no
20		2025-12-03	taxable		1			no
21		2025-12-03	taxable		1			no
22		2025-12-03	taxable		1			no
23		2025-12-03	taxable		1			no
24		2025-12-03	taxable		1			no
25		2025-12-03	taxable		1			no
26		2025-12-03	taxable		1			no
27		2025-12-03	taxable		1			no
28		2025-12-03	taxable		1			no
29		2025-12-03	taxable		1			no
30		2025-12-03	taxable		1			no

	column 17	column 18	column 19	column 20	column 21	column 22	column 23	
1	Backorders allowed?	Sold individually?	Weight(kg)	Length(cm)	Width(cm)	Height(cm)	Allow customer reviews?	
2	no	no					1	
3	no	no					1	
4	no	no					1	
5	no	no					1	
6	no	no					1	
7	no	no					1	
8	no	no					1	
9	no	no					1	
10	no	no					1	
11	no	no					1	
12	no	no					1	
13	no	no					1	
14	no	no					1	
15	no	no					1	
16	no	no					1	
17	no	no					1	
18	no	no					1	
19	no	no					1	
20	no	no					1	
21	no	no					1	
22	no	no					1	
23	no	no					1	
24	no	no					1	
25	no	no					1	
26	no	no					1	
27	no	no					1	
28	no	no					1	
29	no	no					1	
30	no	no					1	

	column 24	column 25	column 26	column 27	column 28	column 29	
1	Purchase note	Sale price	Regular price	Categories	Tags	Shipping class	Images
2		295000	295000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
3		11900	11900				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
4							
5		320000	718000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
6		199000	199000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
7		22000	33000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
8		89000	171000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
9		45000	84000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
10		65000	65000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
11		75000	89000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
12		366000	650000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
13		304000	500000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
14		269000	310000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
15		29000	29000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
16		548000	700000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
17		149000	390000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
18		436000	628000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
19		326000	750000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
20		18000	18000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
21		149000	190000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
22		299000	559000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
23		199000	387000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
24		101000	134500				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
25		84000	84000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
26		398000	599250				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
27		159000	159000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
28		5000	10000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
29		339000	622000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>
30		162000	199000				<a href="https://salt.tikicdn.com/cache/750">https://salt.tikicdn.com/cache/750</a>

	column 31	column 32	column 33	column 34	column 35	column 36	column 37	cc
1	Download limit	Download expiry days	Parent	Grouped products	Upsells	Cross-sells	External URL	Buttc
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								
30								

	2	column 33	column 34	column 35	column 36	column 37	column 38	column 39
1	days	Parent	Grouped products	Upsells	Cross-sells	External URL	Button text	Position
2								0
3								0
4								0
5								0
6								0
7								0
8								0
9								0
10								0
11								0
12								0
13								0
14								0
15								0
16								0
17								0
18								0
19								0
20								0
21								0
22								0
23								0
24								0
25								0
26								0
27								0
28								0
29								0
30								0