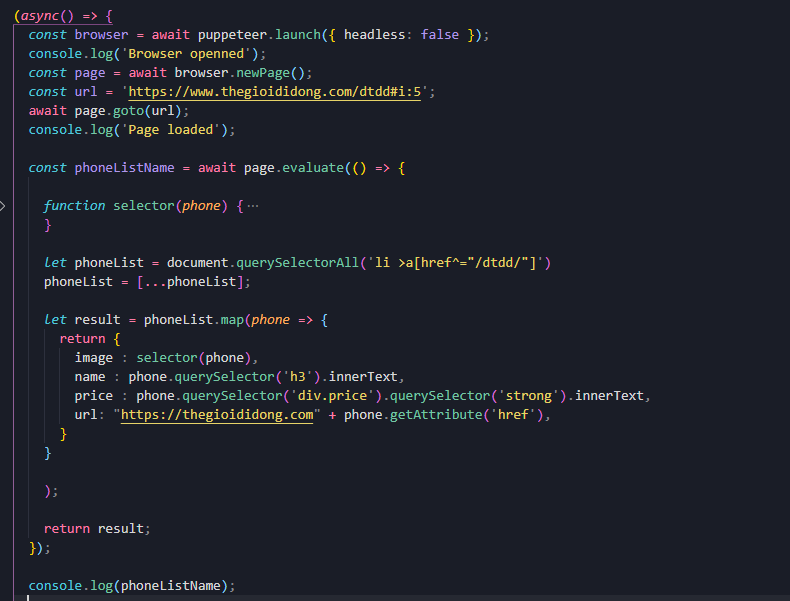
Mục lục

[**1.** **Cào dữ liệu các mẫu điện thoại trên thegioididong bằng Puppeteer:** 1](#_Toc17352748)

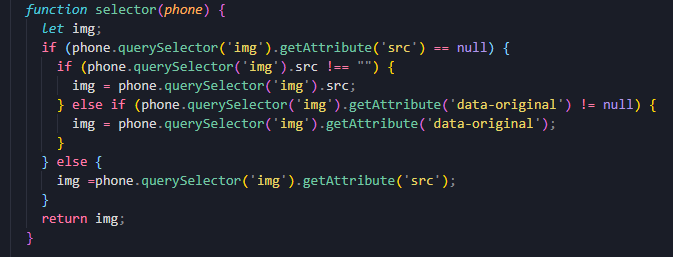
[**2.** **Khảo sát các loại câu hỏi ở trang review mỗi điện thoại:** 4](#_Toc17352749)

# **Cào dữ liệu các mẫu điện thoại trên thegioididong bằng Puppeteer:**

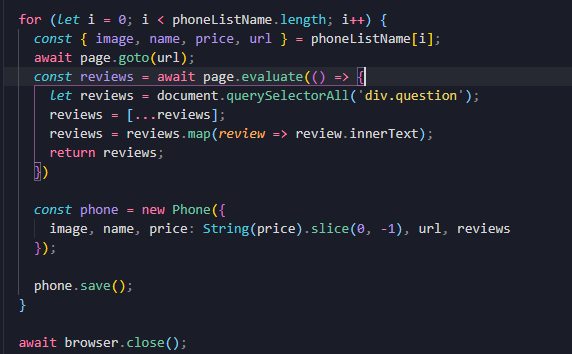
* Puppeteer là một thư viện của NodeJS, hỗ trợ người dung có thể mô phỏng các thao tác cũng như chạy script trên trình duyệt.
* Chi tiết code:



Các hàm như **launch, newPage, goto** giúp ta giả lập mở trình duyệt. Hàm **evaluate()** giúp ta thao tác trên HTML của trình duyệt vừa mở để query đúng dữ liệu ta cần.

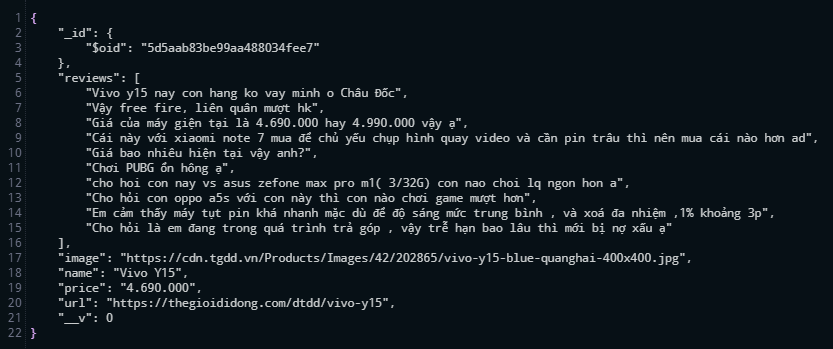


Hàm selector giúp chọn đúng url ảnh của điện thoại, vì trang thegioididong lưu url ảnh dưới nhiều attribute như *src, data-original* nên cần phải chọn lại



Hàm thực hiện chức năng quét qua danh sách điện thoại đã query, mở url của điện thoại đó sau đó query các review về điện thoại để lưu lại. Hàm **phone.save()** insert thông tin đầy đủ của 1 mẫu điện thoại vào database.

Dữ liệu được lưu vào MongoDB, là một cơ sở dữ liệu không quan hệ (NoSQL), được host online trên mlab.com

* **Kết quả thu được**:
* Dữ liệu của 148 mẫu điện thoại trên trang thegioididong, chứa tên, hình ảnh, giá, và 10 bình luận về mẫu điện thoại đó



# **Khảo sát các loại câu hỏi ở trang review mỗi điện thoại:**

* Câu hỏi dạng what:

*"Giá bao nhiêu hiện tại vậy anh?"*

* Cần biết giá của máy. Loại câu hỏi này có thể xác định được entity (“giá”, “tên máy”) rồi thực hiện truy vấn database trả về kết quả.
* Câu hỏi cần tư vấn, so sánh để hỗ trợ ra quyết định:

*"Cái này với xiaomi note 7 mua để chủ yếu chụp hình quay video và cần pin trâu thì nên mua cái nào hơn ad"*

*"cho hoi con nay vs asus zefone max pro m1( 3/32G) con nao choi lq ngon hon a"*

* Cần so sánh chức năng của 2 điện thoại. Loại câu hỏi này có thể được xác định bằng từ khóa “nên”, “cái nào”. Cần so sánh đúng các thuộc tính giữa 2 điện thoại người dung cung cấp và trả về kết quả.
* Câu hỏi dạng why:

*"sao đt em nó cứ quảng cáo hoài vậy ad hễ bật mạng là quãng cáo"*

* Cần xác định nguyên nhân “*cứ quảng cáo hoài vậy*”, rồi đưa ra trả lời tương ứng. Thường những câu hỏi dạng này sẽ liên quan về kĩ thuật.
* Câu hỏi dạng how:

*"Làm sao mo hiển thị thông báo vậy qtv"*

* Cần xác định ý định người dung, thường là sau từ khóa “Làm sao, như thế nào,..” rồi đưa ra câu trả lời tương ứng
* Câu hỏi yes/no:

*"Máy này có khuyến mãi gì không"*

* Xác định bởi từ khóa “có không” ở cuối câu, cần phản hồi lại bằng có/ không, đồng thời cần cung cấp them thông tin sau phản hồi “có/không”
* Một số các khó khan khi phân tích input:

+ Input là chữ tiếng việt không có dấu

+ Input là chữ viết tắt, teencode. Ví dụ: “bjo = bây giờ”

+ Input chứa các từ viết sai chính tả. Ví dụ: “Cho hoi sao minh goi **viedeo** o ben **merrger** no toan tu dong tat k vay”. Trong trường hợp này video và messenger bị viết sai.