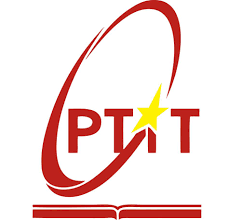
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

**KHOA CÔNG NGHỆ THÔNG TIN 1**



**BÁO CÁO BÀI TẬP LỚN PYTHON**

**PYTHON SCRAPING**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Nhóm học phần:*** | ***Nhóm 5*** | | |
| ***Nhóm bài tập lớn:*** | ***Nhóm 16*** | | |
| ***Sinh viên thực hiện:*** |  |  | |
|  | ***Trần Quang Huy :***  ***(Trưởng nhóm)*** | | ***B20DCCN327*** |
|  | ***Nguyễn Cảnh Hưng:*** | | ***B20DCCN*** |
|  | ***Nguyễn Tiến Anh:*** | | ***B20DCCN*** |
|  | ***Hiếu:*** | | ***B20DCCN*** |
| ***Giảng viên giảng dạy:*** | ***Thầy Nguyễn Hoàng Anh*** | | |

**HÀ NỘI-202**

1. **Giới thiệu sơ lược về sản phẩm của nhóm**
2. **Mô tả ứng dụng**

* Tên ứng dụng: Python scraping
* Nhận thấy được người dùng đang dành nhiều thời gian cho việc mua sắm online, (Cụ thể là sàn TMĐT Shopee), họ có thể dành cả tiếng đồng hồ để phân vân cuối cùng mới quyết định chọn ra sản phẩm cần mua. Quyết định của họ thường dựa trên một số yếu tố nhất định: Shop bán đồ đó có uy tín hay không, số lượng sản phẩm đã bán ra có nhiều hay không và tỷ lệ yêu thích sản phẩm đó như thế nào? … Nhận thấy được điều đó, nhóm 16 có triển khai thực hiện 1 ứng dụng giúp việc mua sắm trở nên dễ dàng hơn. Bằng việc lấy toàn bộ dữ liệu cơ bản về sản phẩm từ ngay trang 1 của shopee.vn. Nhóm 16 đã lọc ra những thông tin cần thiết đủ để khách hàng có thể đưa ra quyết định nên chọn sản phẩm nào.
* Thêm vào đó, nhóm cũng bổ sung tính năng giúp người dùng có thể lấy số lượng ảnh theo yêu cầu các link từ 1 trang web cho trước.
* Phân tích tổng quan ứng dụng:
* Phân tích thiết kế chi tiết ứng dụng:

1. Sơ đồ khối chức năng ứng dụng

Python scraping

Get danh sách sản phẩm từ shopee

Get các ảnh từ 1 trang web bất kì

Lưu và hiển thị các ảnh đã lưu trong thư mục

Sắp xếp danh sách sản phẩm

Lưu danh sách sản phẩm vào file CSV

Hiển thị danh sách sản phẩm lên màn hình

Sắp xếp theo đánh giá sản phẩm

Sắp xếp theo giá giảm dần

Sắp xếp theo doanh số giảm dần

1. Danh sách các lớp đã xây dựng

* Lớp **product** để mô tả các sản phẩm lấy từ shopee.vn

Bao gồm: Tên, giá nhỏ nhất, giá lớn nhất, đánh giá sản phẩm, tổng số lượt bán ra của shop đó và cuối cùng là link dẫn đến sản phẩm đó trên shopee

Biểu đồ:

fdProduct

* Name: str
* Min price: int
* Max price: int
* Rating: float
* Sales: int
* Link: str
* \_\_str\_\_()
* \_\_inter\_\_()
* Lớp initWindow kế thừa từ tkinter.Tk(): Để khởi tạo GUI

initWindow(tk.Tk)

* Tk: tkinter.Tk()
* \_\_init\_\_()
* Intro()
* initPage()
* run()
* Ngoài ra, nhóm 16 còn tạo ra 2 file py chứa các hàm được sử dụng trong chương trình.
  + File1: shopeeFunctions.py: chứa các hàm liên quan đến Chức năng thứ nhất: Lấy danh sách sản phẩm từ shopee theo tên tìm kiếm

shopeeFunctions.py

* accessToGithub()
* getPosition()
* generateLink()
* getHtml()
* showProgressBar()
* endProgress()
* fillProductList()
* writeToFile()
* clearTreeView()
* fillTreeView()
* showProducts()
* accessToShopee()
  + File2: ImageFunctions.py chứa các hàm liên quan đến Chức năng thứ hai: Lấy các ảnh từ 1 website rồi lưu vào thư mục đã được định sẵn

imageFunctions.py

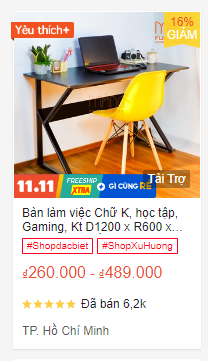
* Get\_url()
* showProgressBar()
* endProgressBar()
* download()
* getImage\_run()
* showFolder()

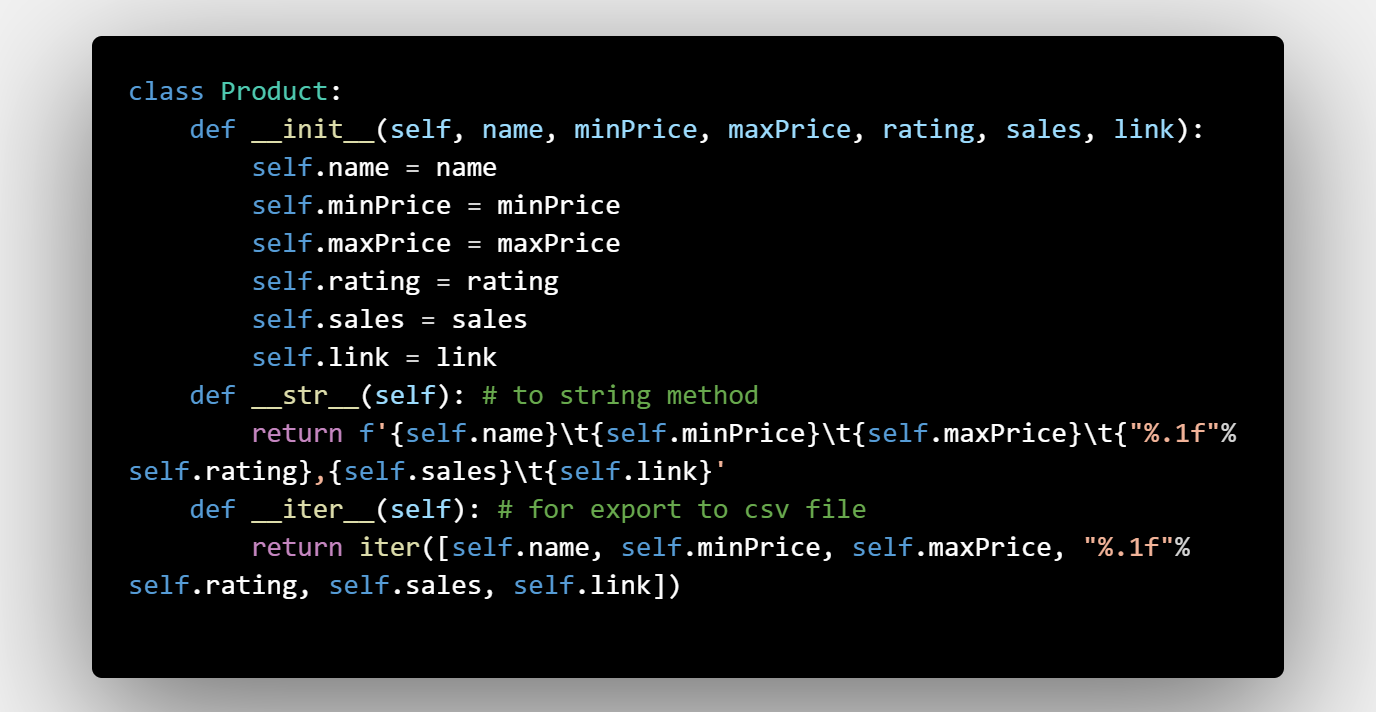
1. Sơ đồ thực thể quan hệ ER
2. **So sánh nội dung với github được giao**

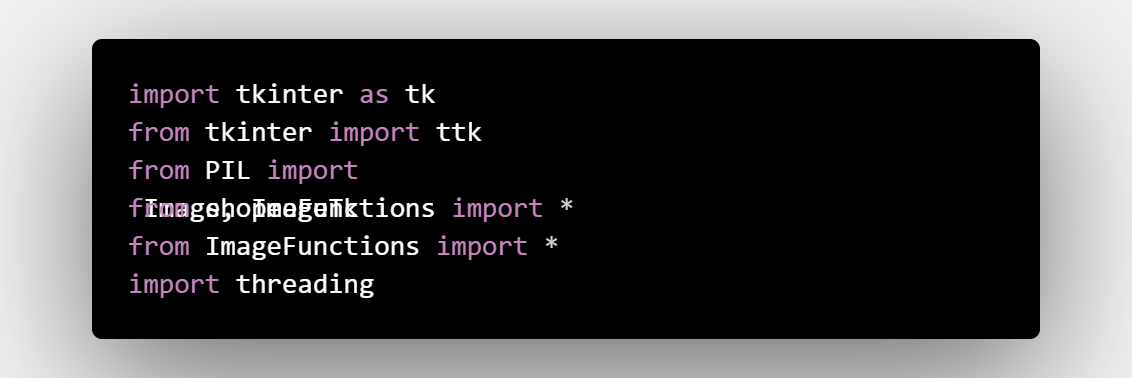
* Sau khi tìm hiểu về git: <https://github.com/REMitchell/python-scraping.git>. Team 16 nhận thấy danh sách source code được push lên là code hướng dẫn của cuốn sách cùng tên của tác giả Ryan Mitchell. Sách gồm 18 chương, mỗi chương là 1 phần hướng dẫn về cách crawl dữ liệu từ web. Đi cùng với đó là code sample demo cho từng phần.
* Vì lí do đó team 16 quyết định sẽ phát triển mới 1 ứng dụng để crawl dữ liệu có vận dụng kiến thức đã được đề cập trong sách.

1. **Kết quả ứng dụng**
2. **Giải thích code cho từng phần**

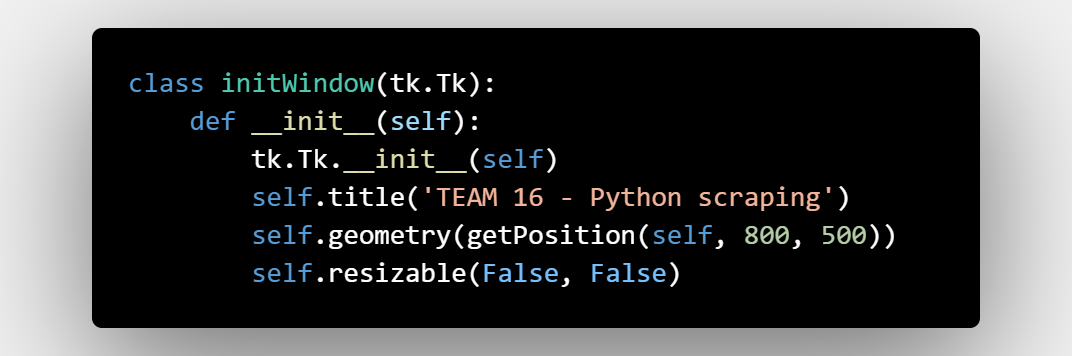
* **Hàm chứa lớp product: Product.py**
* Từ phân tích sơ lược trang web shopee.vn, nhóm 16 đã rút ra được 5 thuộc tính cơ bản mà người dùng có thể quan tâm khi mua sắm trên web.

Từ 1 sản phẩm ngẫu nhiên: ta có thể thấy được các thông tin cơ bản như tên sản phẩm, giá nhỏ nhất, giá lớn nhất, lượt đánh giá của sản phẩm (sao), doanh số sản phẩm và 1 thuộc tính ẩn là đường link dẫn đến sản phẩm đó  
- Từ đó, nhóm xây dựng lớp product như sau

Hình 0: Xây dựng lớp product

* Với 2 phương thức: \_\_str\_\_() dùng cho việc ghi sản phẩm ra màn hình console, \_\_íter\_\_() dùng cho việc ghi sản phẩm vào file csv.
* **Hàm khởi tạo GUI: init.py**
* Các thư viện đã sử dụng: tkinter bao gồm tk và ttk, PIL, threading
* Các module tự xây dựng: shopeeFunctions và imageFunctions

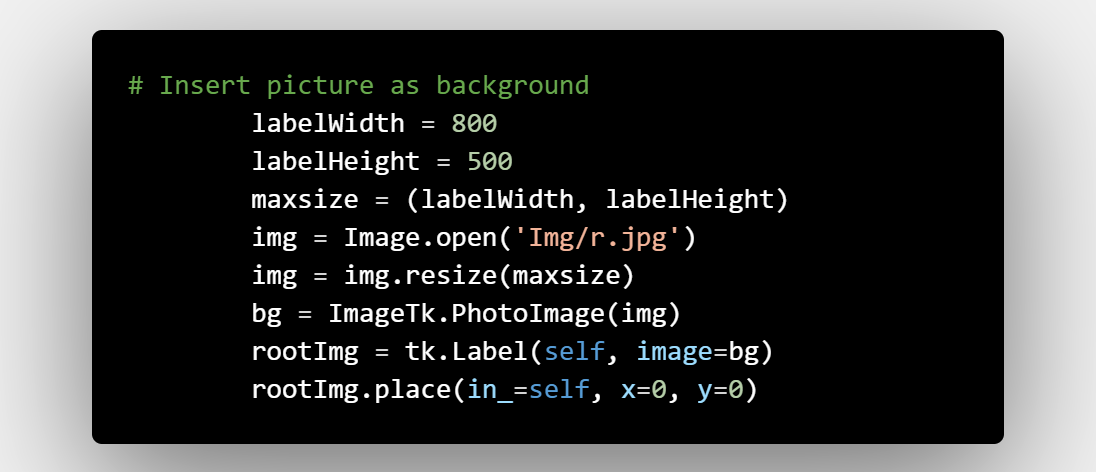
Hình 1. Các thư viện sử dụng trong init.py

* Nhóm đã xây dựng lớp initWindow kế thừa từ tk.Tk() với hàm khởi tạo chỉnh các yếu tố về vị trí và kích thước cho lớp:

Hình 2. Xây dựng lớp initWindow

* Trong lớp initWindow, nhóm đã xây dựng hàm initPage() với mục đích thêm các widget con vào trong Tk() frame ban đầu.

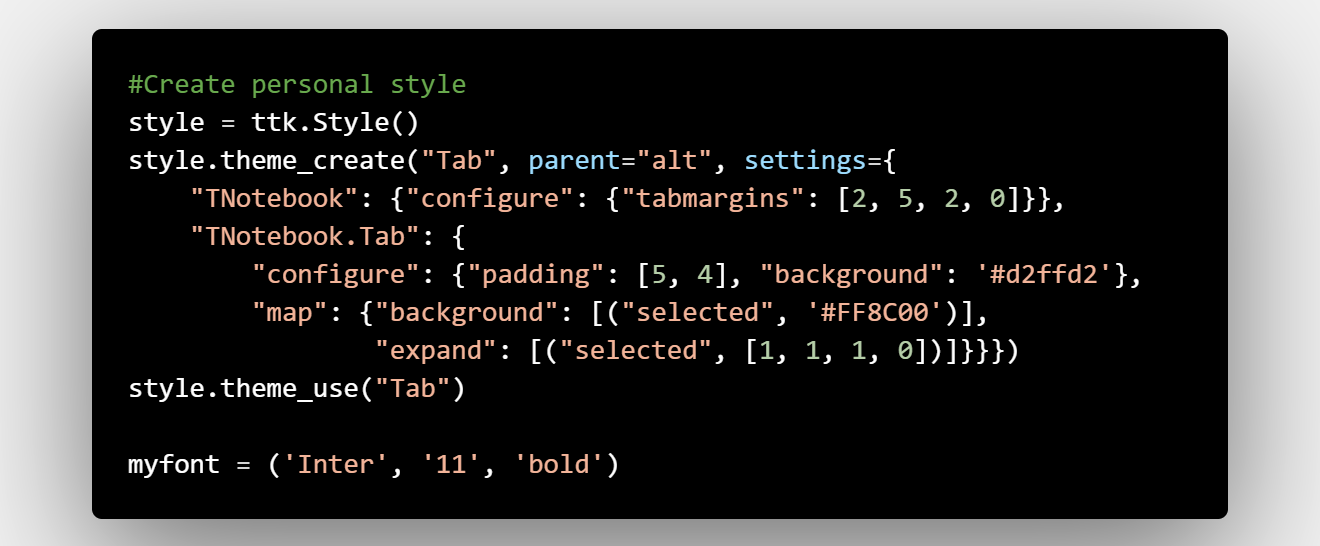
Hình 3. Hàm thêm các widget con vào Tk()

* Các widget con bao gồm:
  + - Image, ImageTk, label dùng để chèn background

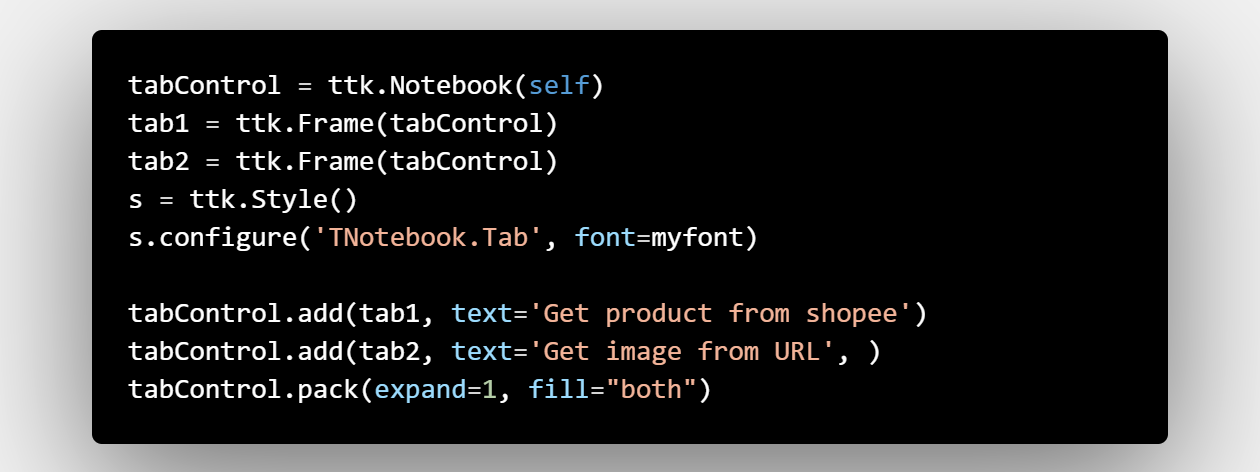
Hình 4. Khởi tạo background

* + Button: button về link github của nhóm

Hình 5. Button về link github

* + Tạo style riêng và font chữ dùng cho ứng dụng

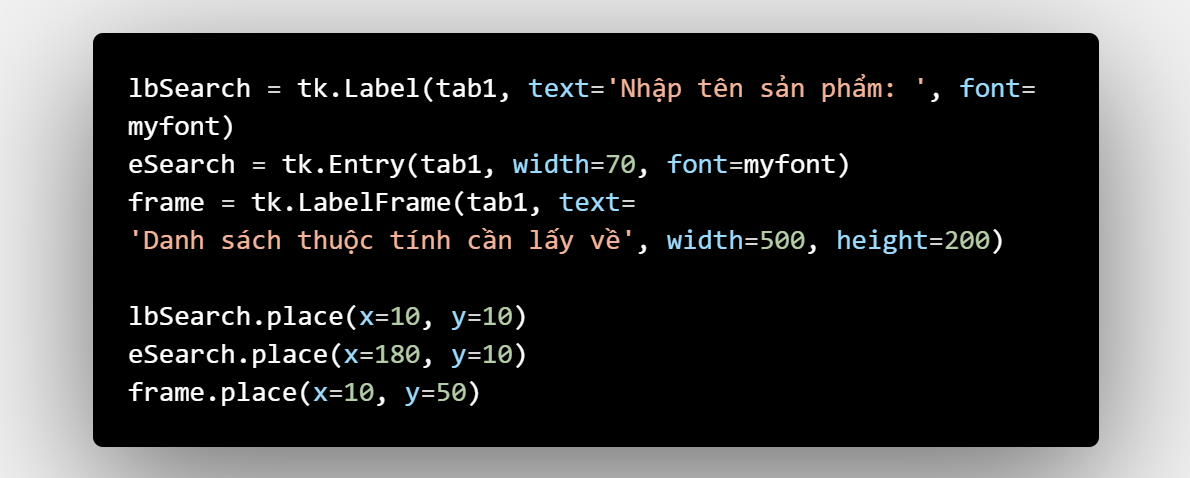
Hình 6. Tạo style và font

* + Xây dựng widget tab và sử dụng style và font đã cài bên trên cho tab

Hình 7. Tab

Sau khi chia thành 2 tab riêng biệt, nhóm 16 tiến hành xây dựng giao diện cho tab thứ 1:

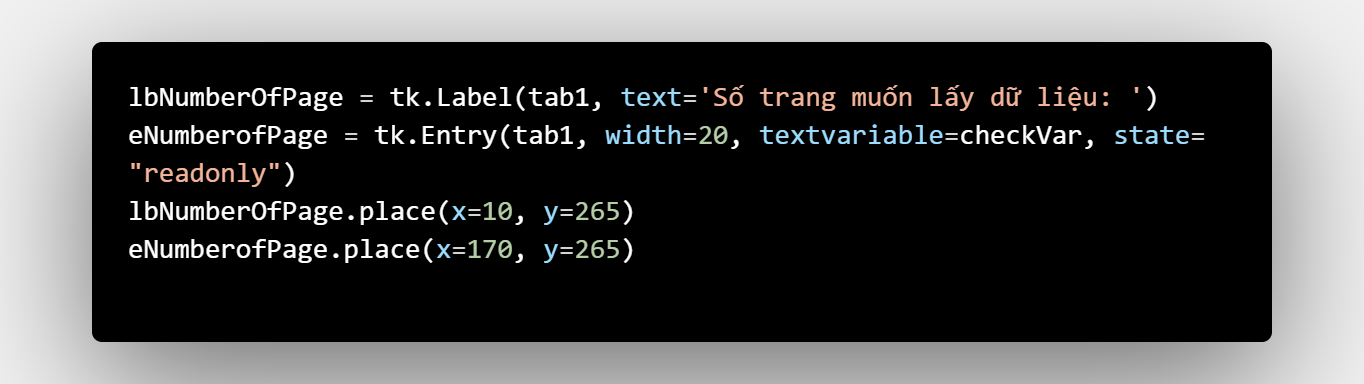
* + Các label dùng để hướng dẫn người dùng

Hình 8: label, entry và Frame hướng dẫn người dùng

* + Các checkbox, phần này dùng để phát triển ứng dụng trong tương lai

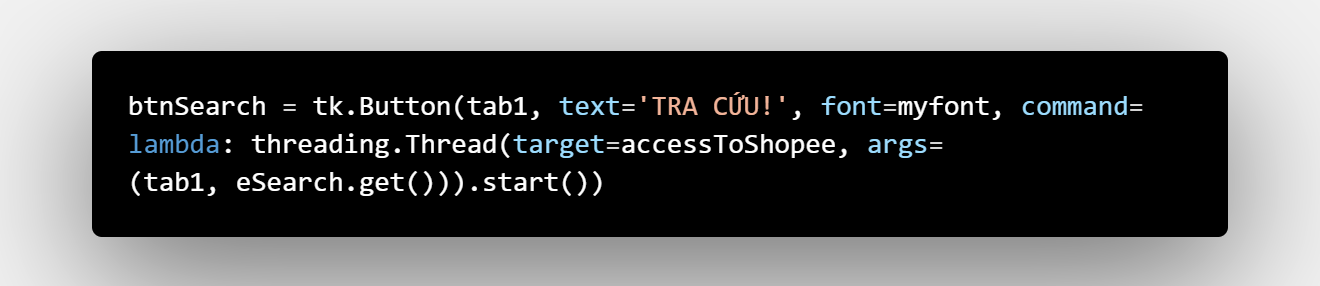
Hình 9: các checkbox

* + Label hiển thị số trang muốn lấy dữ liệu và entry phần nhập (Phần này cũng để phát triển ứng dụng trong tương lai)

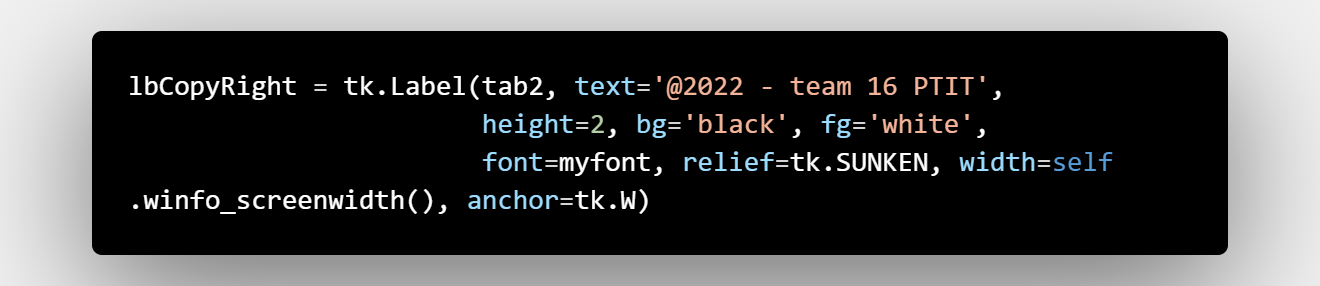
Hình 10. Label và Entry số trang

* + - Logo của shopee

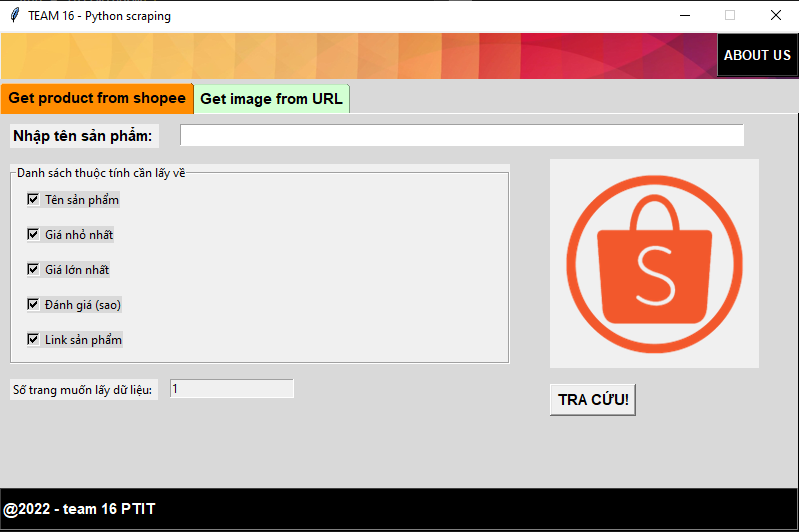
Hình 11. Logo shopee

* + - Button dùng để tra cứu

Hình 12. Button tra cứu

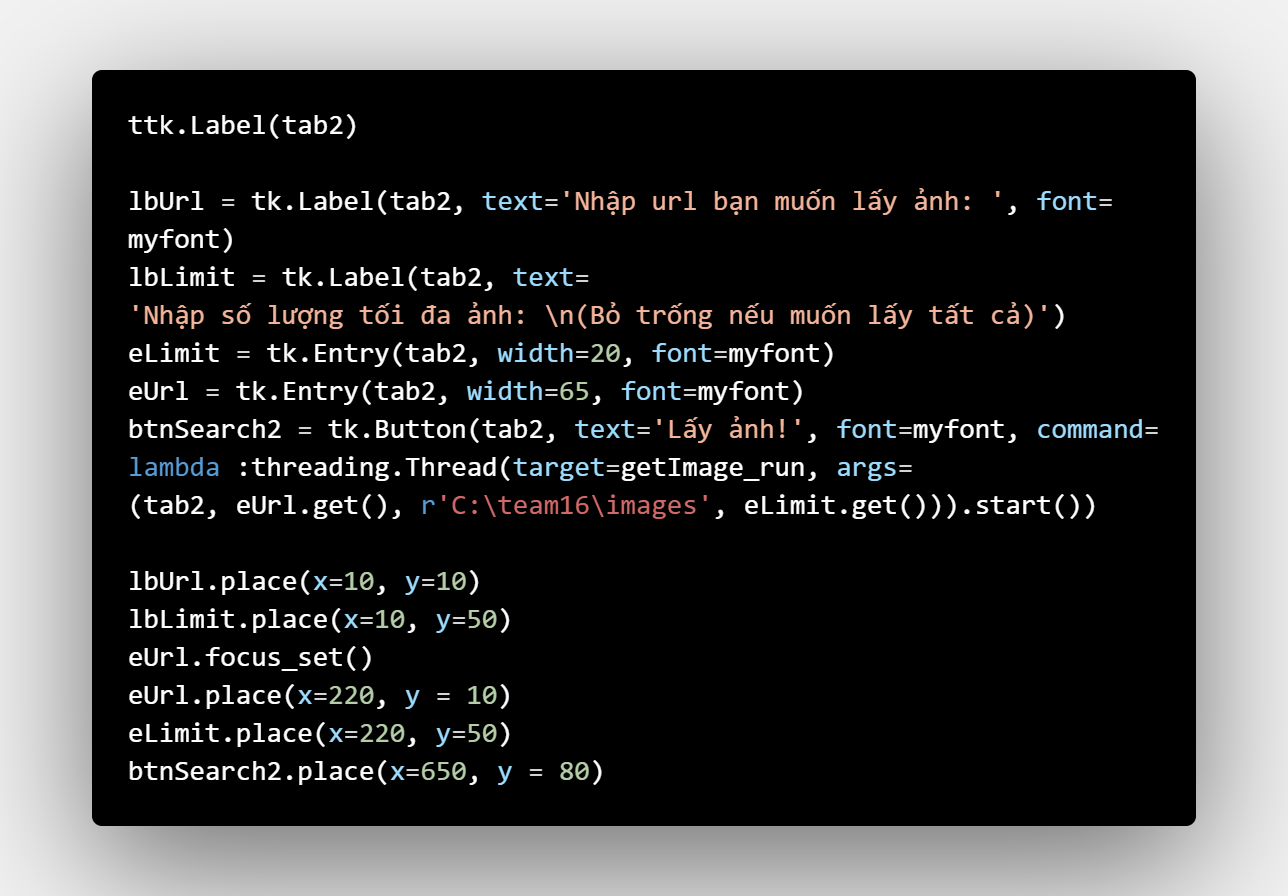
* + Label copy right

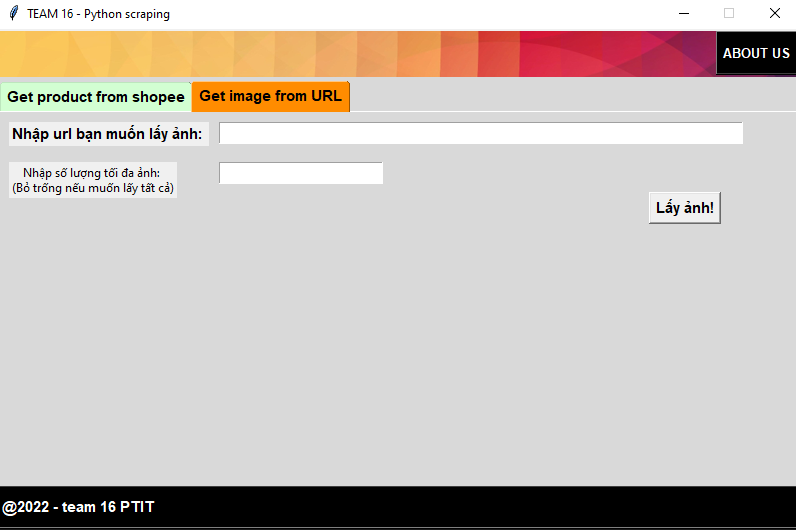
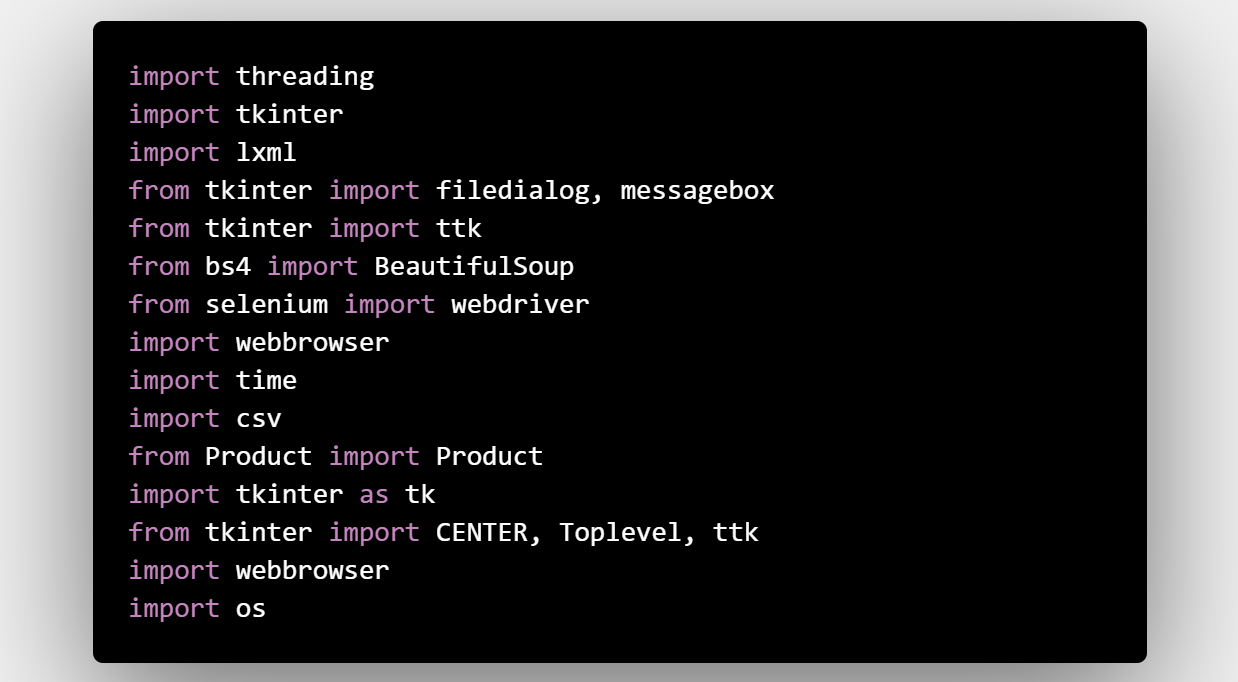
Hình 13. Label Copyright

* Sau khi hoàn thiện, GUI của tab 1 sẽ như sau

Hình 14. GUI của tab thứ nhất

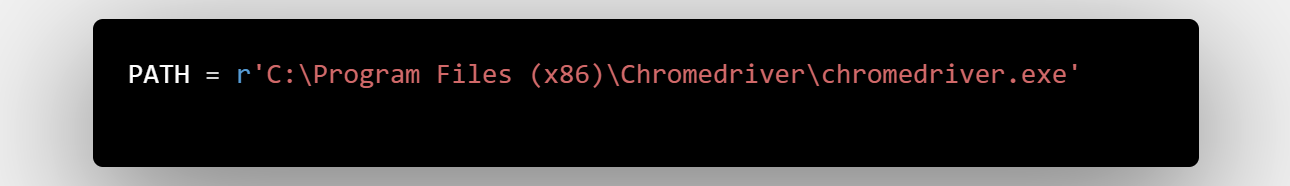
* Tab thứ 2 được xây dựng đơn giản hơn với 2 label và 1 entry

Hình 15. Tab2

* Sau khi xây dựng, tab2 sẽ có giao diện như sau
* Sau khi đã xây dựng xong GUI, ta gọi hàm mainloop() của Tk() để hiển thị lên màn hình
* Hàm để chạy phương thức initPage()
* **Module shopeeFunctions.py dùng để xây dựng các hàm lấy về dữ liệu từ shopee**
* Các thư viện và module dùng để lấy data về bao gồm

Hình 17. Các thư viện dùng trong shopeeFunctions.py

* Vì app có dùng thư viện selenium, do đó cần cài Chromedriver từ google. Sau khi tải xong thì copy đường dẫn và gán nó vào biến PATH

Hình 18. Đường dẫn tới Chromedriver

* Vì ở class initWindow, nhóm có xây dựng button About Us, khi nhấn nó sẽ dẫn tới link source code của nhóm trên github. Trong đó, hàm dẫn tới github của nhóm được xây dựng riêng trong

1. **Cài đặt triển khai ứng dụng**
2. **Kết luận, hạn chế và phương hướng phát triển trong tương lai**
3. **Tài liệu tham khảo**