**BỘ MÔN HỆ THỐNG THÔNG TIN – KHOA CÔNG NGHỆ THÔNG TIN**

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM**

Thực hiện: Nhóm 8 – lớp 17\_1

GV phụ trách: Lê Nguyễn Hoài Nam

ĐỒ ÁN môn học  - CHUYÊN ĐỀ CHỌN LỌC HỆ THỐNG THÔNG TIN

HỌC KỲ I – NĂM HỌC 2020-2021

**môn CHUYÊN ĐỀ CHỌN LỌC TRONG HỆ THỐNG THÔNG TIN**

**BẢNG THÔNG TIN CHI TIẾT NHÓM**

|  |  |  |
| --- | --- | --- |
| **Mã nhóm:** | Nhóm 8 | |
| **Số lượng:** | **5** | |
| **MSSV** | **Họ tên** | **Công việc** |
| 1712120 | Trần Thái Như | Thiết kế CSDL, import data |
| 1712169 | Bùi Phan Thọ | Tìm hiểu các api, thư viện python cần thiết |
| 1712388 | Nguyễn Phạm Nhật Duy | Video thuyết minh, báo cáo file word |
| 1712629 | Nguyễn Nhân | Code |
| 1712647 | Nguyễn Văn Núi | Code |

**YÊU CẦU ĐỒ ÁN- BÀI TẬP**

|  |  |
| --- | --- |
| **Loại bài tập** | **🗹 Lý thuyết Thực hành 🗹 Đồ án Bài tập** |
| **Ngày bắt đầu** | **1/12/2020** |
| **Ngày kết thúc** | **30/1/2021** |

# **Yêu cầu của Đồ án**

* Thực hành crawl dữ liệu từ Tiki
* Import dữ liệu vào Hệ quản trị CSDL bất kỳ
* Source code crawl dữ liệu
* Video demo
* Báo cáo các kết quả đạt được

# **Kết quả**

## **Mục lục**

[A. Yêu cầu của Đồ án 2](#_Toc62410465)

[B. Kết quả 3](#_Toc62410466)

[Mục lục 3](#_Toc62410467)

[Nội dung chi tiết 4](#_Toc62410468)

[I/ Ý tưởng thực hiện: 4](#_Toc62410469)

[II/ Mô tả source code đồ án: 5](#_Toc62410470)

[**1.** **Môi trường lập trình:** 5](#_Toc62410471)

[**2.** **Thư viện sử dụng:** 5](#_Toc62410472)

[**3.** **Các bước lập trình:** 5](#_Toc62410473)

[*a.* *Cài đặt các thư viện cần thiết:* 5](#_Toc62410474)

[*b.* *Giải thích ý tưởng và các hàm thực thi tương ứng:* 5](#_Toc62410475)

[**4.** **Kết quả:** 16](#_Toc62410476)

[**5.** **Thiết kế CSDL:** 19](#_Toc62410477)

[*a.* *Mô hình CSDL:* 19](#_Toc62410478)

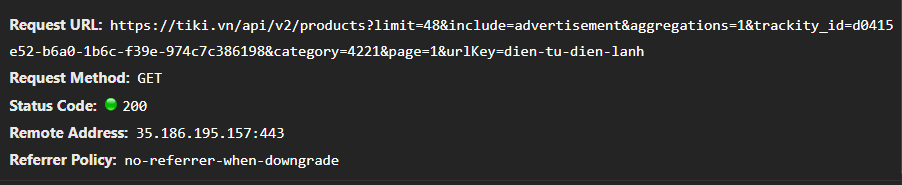
[*b.* *Hướng dẫn import dữ liệu bằng SQL Server Wizard:* 21](#_Toc62410479)

[**6.** **Hướng dẫn sử dụng source code:** 26](#_Toc62410480)

## **Nội dung chi tiết**

## **I/ Ý tưởng thực hiện:**

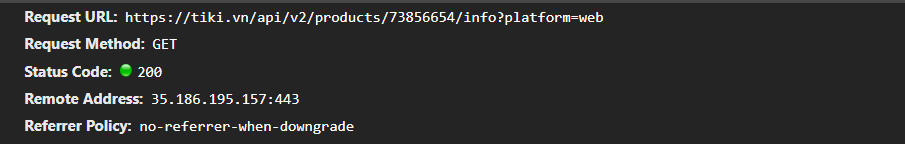
* Thử inspect trang Tiki trong danh mục sản phẩm Điện tử - Điện lạnh: <https://tiki.vn/dien-tu-dien-lanh/c4221>
* Vào tab Network, Ctrl + F5 và tìm kiếm các request XHR
* Ta thấy được 1 request như sau:



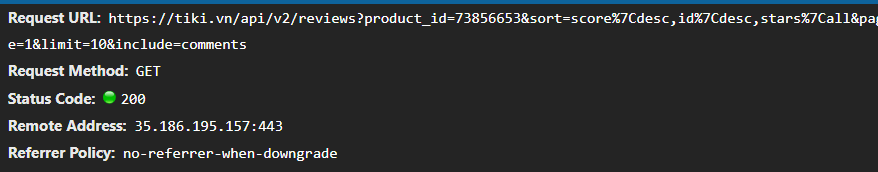
* Thử mở trong tab khác ta được 1 chuỗi json:



* Để ý kỹ ta sẽ thấy tên sản phẩm cũng như một số thông tin khác, ngoài ra api vừa tìm được có cả page = 1 nên khá chắc đây là api tải 1 trang sản phẩm
* Tiếp theo inspect 1 sản phẩm bất kỳ tìm được api lấy thông tin sản phẩm như sau:



* Tương tự tìm được api lấy bình luận sản phẩm



* Vậy ta cần trích thông tin từ các api tìm được.

## **II/ Mô tả source code đồ án:**

1. **Môi trường lập trình:**

* *Ngôn ngữ lập trình*: Python
* *Môi trường:* Google Colab
* Máy tính có kết nối mạng
* *Hệ quản trị CSDL:* SQL Server

1. **Thư viện sử dụng:**

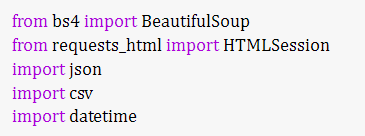
* *Requests-html*: bao gồm các hàm phân tích chuỗi HTML để rút trích thông tin, tạo request lấy dữ liệu từ trang web, …
* *HTMLSession*: là một hàm trong thư viện requests-html dùng để tạo kết nối (phiên làm việc) đến một trang web
* *BeautifulSoup*: gồm các hàm rút trích dữ liệu từ văn bản HTML, …
* *json*: gồm các hàm xử lý chuỗi json, chuyển đổi từ văn bản dạng khác sang json, …
* *csv*: các hàm xử lý đọc, ghi trên file csv
* *datetime*: chuyển đổi format ngày tháng, tính toán ngày tháng, …

1. **Các bước lập trình:**
2. *Cài đặt các thư viện cần thiết:*

* Cài đặt thư viện requests-html bằng lệnh pip



* Import các thư viện cần thiết:

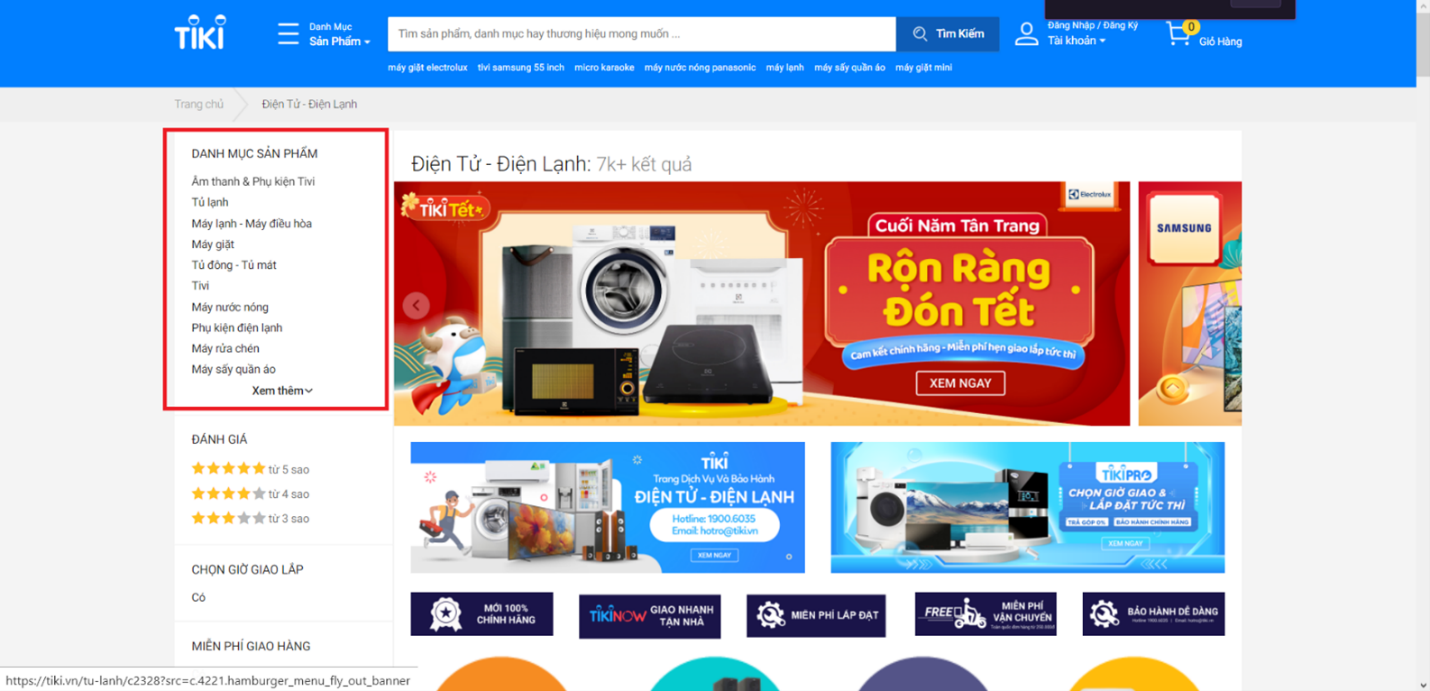


1. *Giải thích ý tưởng và các hàm thực thi tương ứng:*

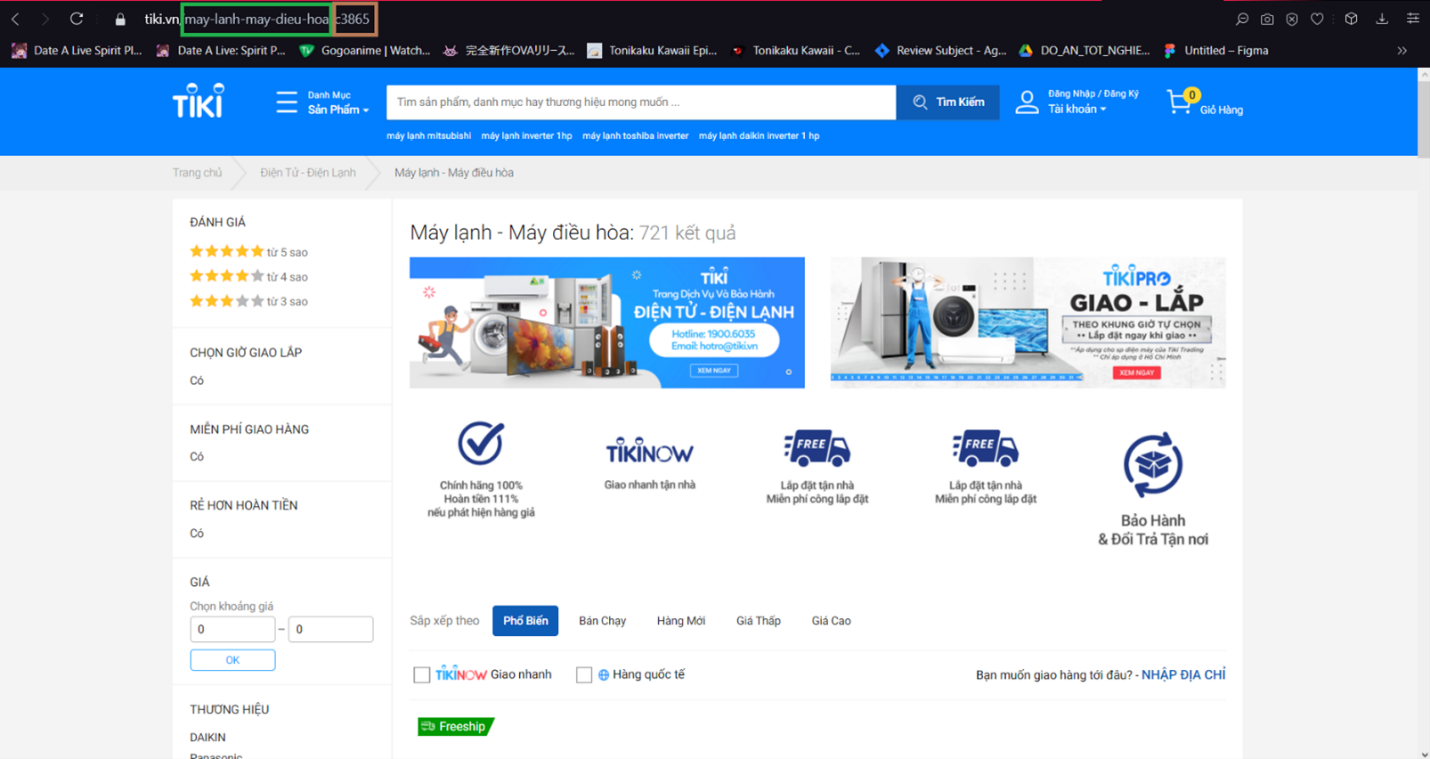
***Ý tưởng:*** gọi api và lấy dữ liệu từ chuỗi html trả về

1. ***Bước 1:***

* Đầu tiên ta tìm các Category ID (mã phân loại sản phẩm) và các url key tương ứng.

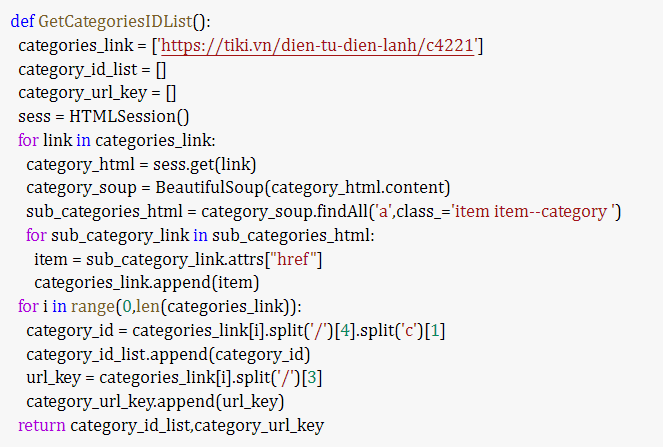


* Với mỗi danh mục sản phẩm sẽ bao gồm nhiều sản phẩm và có một category ID tương ứng nếu bạn click vào và nhìn trên link.



C865 là category ID và may-lanh-may-dieu-hoa là url key

* Do ta cần toàn bộ dữ liệu sản phẩm nên ta phải duyệt từng link danh mục sản phẩm và lấy sản phẩm của danh mục đó về.
* Code:



* Các bước làm của đoạn code trên như sau:

+ Đầu tiên kết nối đến <https://tiki.vn/dien-tu-dien-lanh/c4221> và lấy chuỗi HTML về



+ Tiếp theo ta chuyển chuỗi trả về thành văn bản HTML bằng thư viện BeautifulSoup

Ảnh có chứa văn bản

Mô tả được tạo tự động

+ Ta tìm các thẻ a có class là ‘item item – category’ bằng lệnh findAll của thư viện BeautifulSoup, vì nếu inspect trang web sẽ thấy mỗi thẻ như vậy sẽ chứa thông tin của một danh mục sản phẩm.

Ảnh có chứa văn bản

Mô tả được tạo tự động

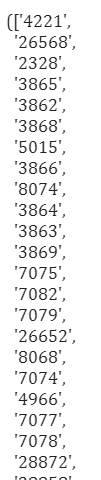
+ Ứng với danh sách các thẻ a vừa tìm được, ta lấy thuộc tính “href” như truy xuất mảng thông thường và thêm vào categories\_link, categories\_link sẽ như sau:



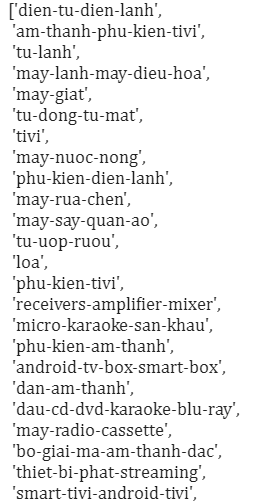
+ Cuối cùng duyệt qua categories\_link, thực hiện tách chuỗi bằng hàm split và thêm vào hai danh sách tương ứng là category\_id và category\_url\_key

* Output:

+ category\_id:



+ category\_url\_key



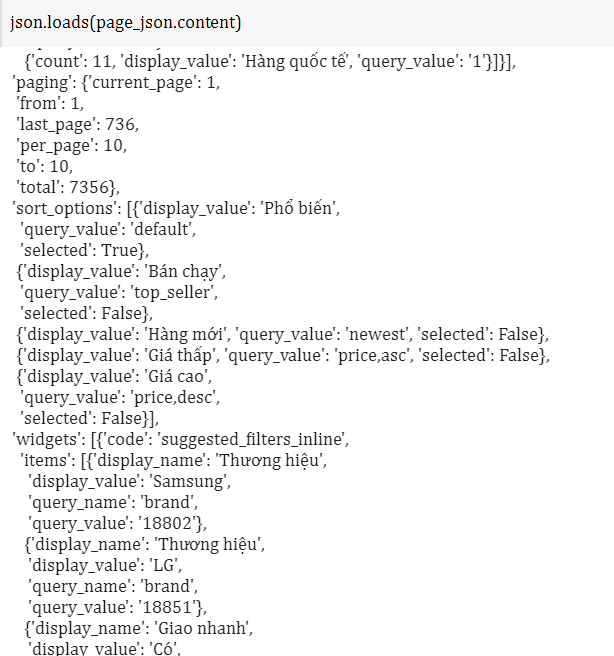
1. ***Bước 2:***

* Sau khi có được danh sách các danh mục sản phẩm ở bước 1, tiếp theo là ta phải tìm các id sản phẩm có trong từng danh mục trong danh sách
* Api sử dụng: [https://tiki.vn/api/v2/products?category={}&page={}&urlKey={}](https://tiki.vn/api/v2/products?category=%7b%7d&page=%7b%7d&urlKey=%7b%7d), ví dụ với [https://tiki.vn/api/v2/products?category=4221&page=1&urlKey=dien-tu-dien-lanh](https://tiki.vn/api/v2/products?category=4221&page=1&urlKey=dien-tu-dien-lanh%7d)

Ảnh có chứa văn bản, ảnh chụp màn hình

Mô tả được tạo tự động

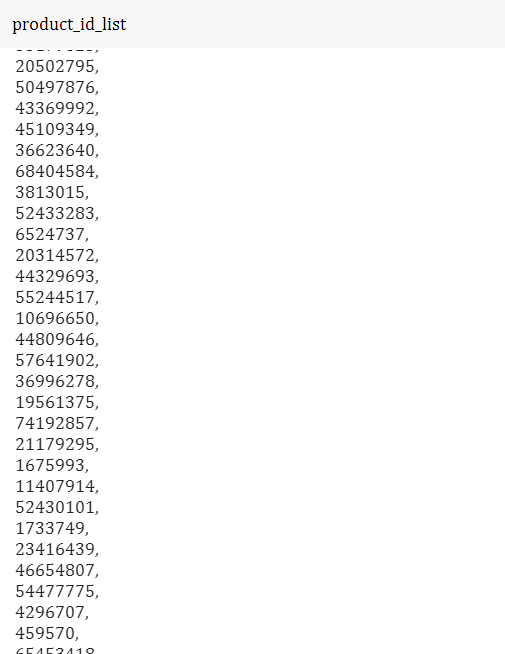
* Ta thấy api trả về kết quả chưa như mong muốn vì có một số chỗ có ký tự Unicode thì bị chuyển đổi thành ký tự khác, vậy chuỗi json này chưa unescaped.
* Ta sử dụng json.loads của thư viện json để chuyển chuỗi json unescaped thành json escaped.



* Để ý kỹ sẽ thấy thông tin các sản phẩm nằm toàn bộ trong thuộc tính ‘data’, vậy cần duyệt hết thuộc tính data thì qua danh mục khác, còn không thì tiếp tục trang số page.
* Ta lấy thuộc tính id sản phẩm với mỗi phần tử có trong thuộc tính data và thêm vào product\_id\_list.
* Code:



* Output: product\_id\_list

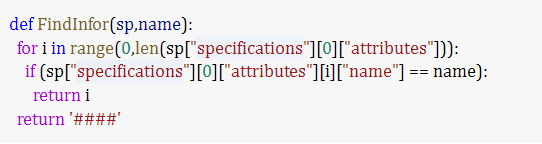


1. ***Bước 3:***

* Sau khi có được toàn bộ id sản phẩm thì công việc tiếp theo sẽ là lấy thông tin của sản phẩm
* Api sử dụng: [https://tiki.vn/api/v2/products/{}](https://tiki.vn/api/v2/products/%7b%7d)
* Ở đây chỉ cần rút trích các thuộc tính mà ta cần, lưu ý cần xử lý một số thuộc tính không có ở một số sản phẩm, …
* Code:

Ảnh có chứa văn bản

Mô tả được tạo tự độngẢnh có chứa văn bản

Mô tả được tạo tự động

* Chú thích các hàm:

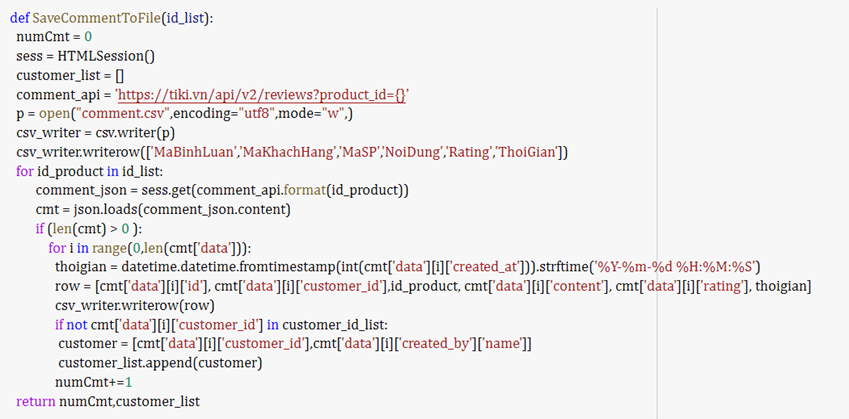
+ Hàm FindInfor dùng để xử lý các thuộc tính bị thiếu của sản phẩm, trả về ‘####’ nếu không tìm thấy, theo như quan sát thì chỉ có 2 trường thường bị thiếu là “Xuất xứ” và “Xuất xứ thương hiệu”

+ Hàm getRow chỉ cần chú ý đến thuộc tính Mô tả của sản phẩm sẽ chứa các thẻ HTML, vì thế ta loại bỏ thẻ bằng cách chuyển thành văn bản HTML với thư viện BeautifulSoup và chỉ lấy text của văn bản

+ Hàm SaveProductToFile sẽ thực hiện vòng lặp lấy thông tin sản phẩm, mỗi lần lấy xong sẽ ghi vào file, nếu thất bại sẽ báo lỗi và thực hiện lấy lại.

1. ***Bước 4:***

* Tương tự bước 3, ta lấy thông tin bình luận của khách hàng
* Api sử dụng: [https://tiki.vn/api/v2/reviews?product\_id={}](https://tiki.vn/api/v2/reviews?product_id=%7b%7d)
* Code:

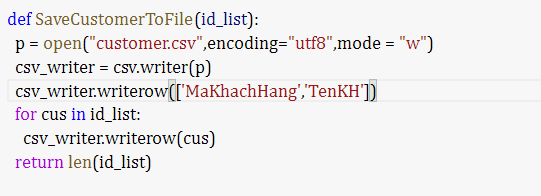


Với hàm này cần chú ý thuộc tính created\_at thể hiện cho thời gian bình luận ở dạng UNIX Time, do đó cần dùng hàm *datetime.fromtimestamp(int()).strftime('%Y-%m-%d %H:%M:%S')* của thư viện datetime để chuyển đổi

Trong đoạn code trên ta cũng lấy được thông tin mã khách hàng và tên của các khách hàng đã bình luận và trả về customer\_list

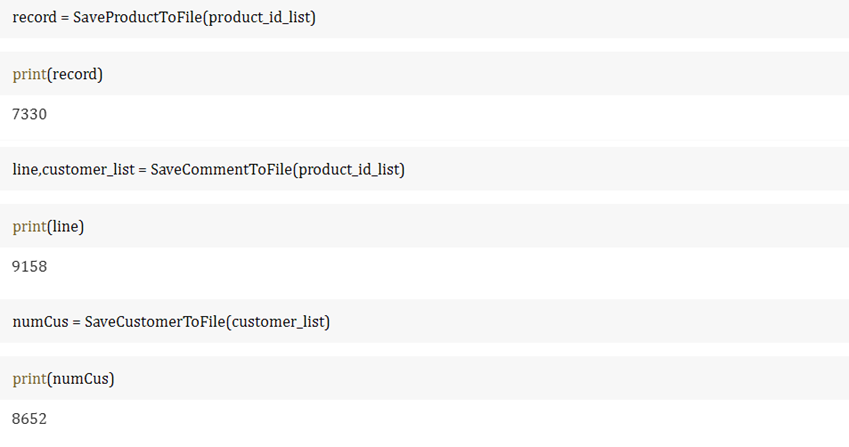
1. ***Bước 5:***

* Với danh sách khách hàng vừa có ở bước 4, ta cài đặt hàm SaveCustomerToFile để ghi danh sách này vào csv và trả về số dòng ghi được
* Code:



1. **Kết quả:**

* Với 3 hàm SaveProductToFile, SaveCommentToFile và SaveCustomerToFile có kết quả trả về là số dòng ghi được vào file csv. Sau khi chạy ta xem số dòng thu được.



* Sau khi chạy code ta thu được 3 file product.csv, comment.csv và customer.csv ở biểu tượng thư mục góc trái

Ảnh có chứa văn bản

Mô tả được tạo tự động

* Tải 3 file về và import bằng excel, cần chỉnh mã là UTF-8.

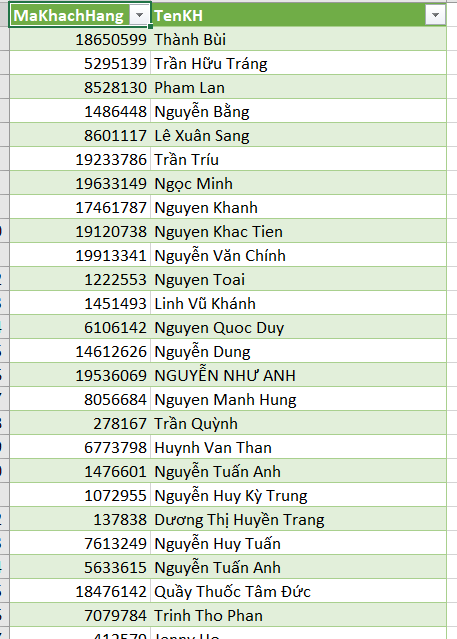
Ảnh có chứa bàn

Mô tả được tạo tự động

* File này sẽ được sử dụng để import dữ liệu vào hệ quản trị SQL server.
* Dữ liệu của 3 bảng:

Ảnh có chứa bàn

Mô tả được tạo tự động

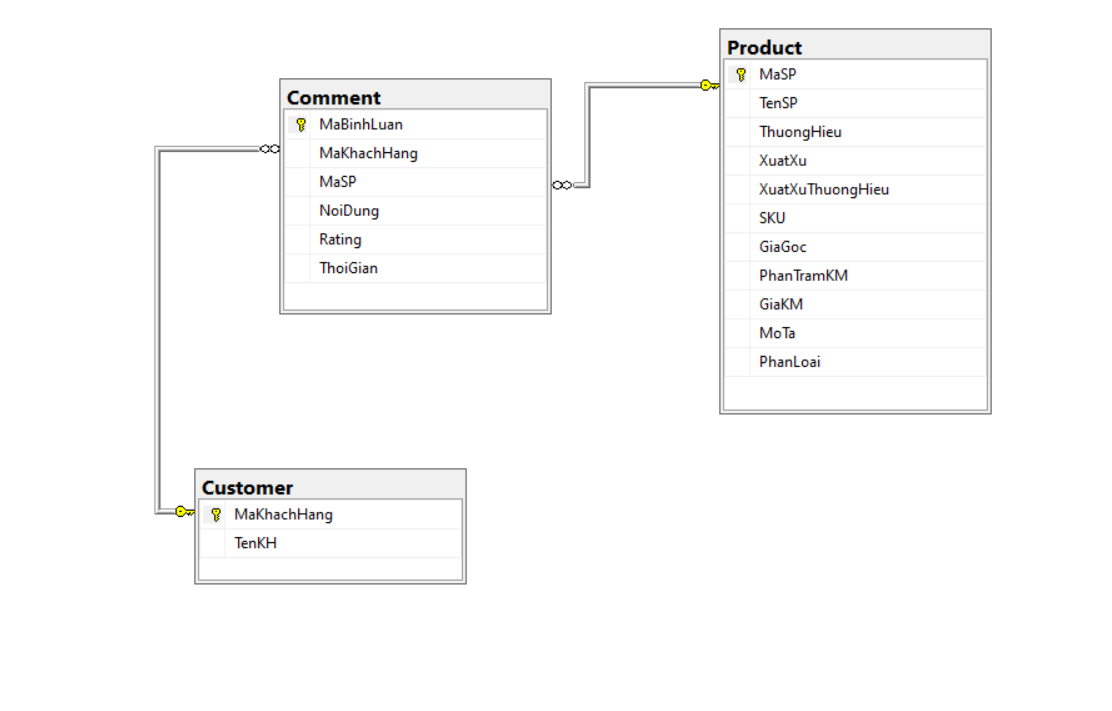


Ảnh có chứa văn bản

Mô tả được tạo tự động

1. **Thiết kế CSDL:**
2. *Mô hình CSDL:*

* Với yêu cầu của đồ án, ta thấy rằng CSDL gồm 3 bảng, đó là bảng Product, Comment, Customer
* Thuộc tính của bảng Product gồm: Mã sp, tên sp, thương hiệu, xuất xứ, xuất xứ thương hiệu, SKU, Giá gốc, giá KM, phần trăm KM, mô tả, phân loại
* Thuộc tính của bảng Comment gồm: Mã bình luận, Mã sp (thể hiện bình luận của sản phẩm nào), Mã khách hàng (thể hiện bình luận của khách hàng nào), Thời gian, Rating, Nội dung
* Thuộc tính của bảng Customer gồm: Mã khách hàng, tên khách hàng
* Sơ đồ thiết kế CSDL:



* Các ràng buộc khóa chính gồm: Product(MaSP), Comment(MaBinhLuan), Customer(MaKhachHang)
* Các ràng buộc khóa ngoại gồm:

+ Một sản phẩm không có hoặc có rất nhiều bình luận

+ Một bình luận chỉ thuộc về một và chỉ một sản phẩm

+ Một bình luận chỉ thuộc về một và chỉ một khách hàng

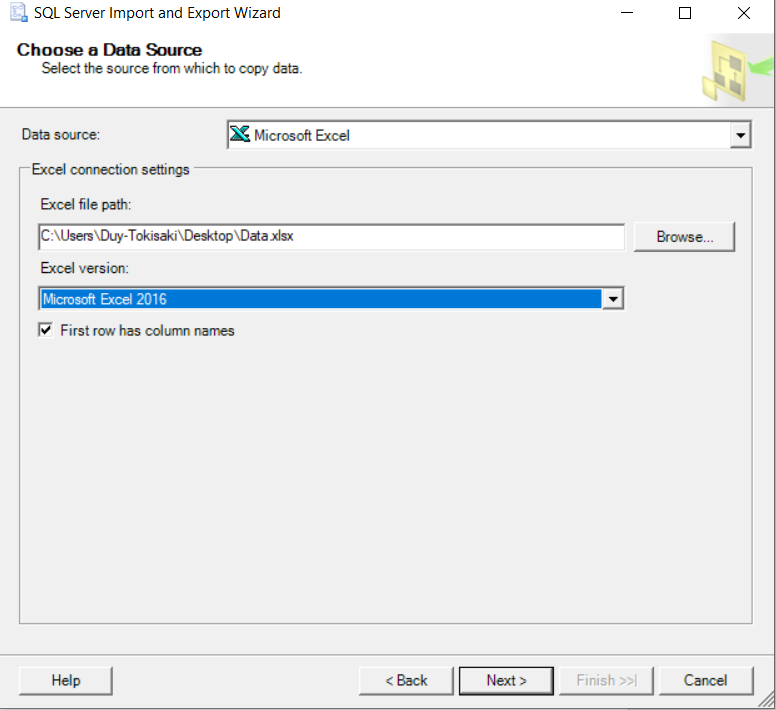
+ Một khách hàng có thể không có hoặc có nhiều bình luận

1. *Hướng dẫn import dữ liệu bằng SQL Server Wizard:*

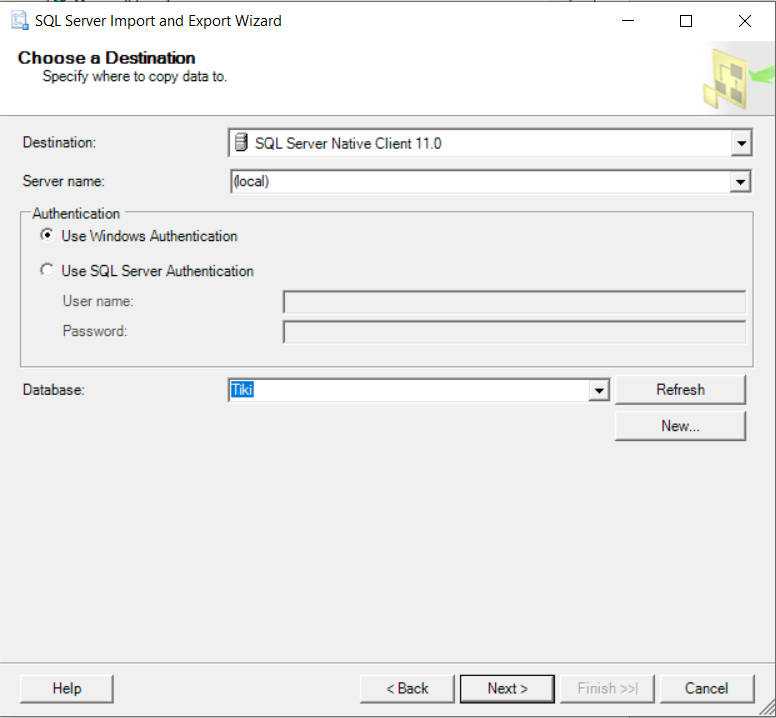
* Tạo CSDL Tiki với lệnh CREATE DATABASE Tiki trên giao diện SSMS



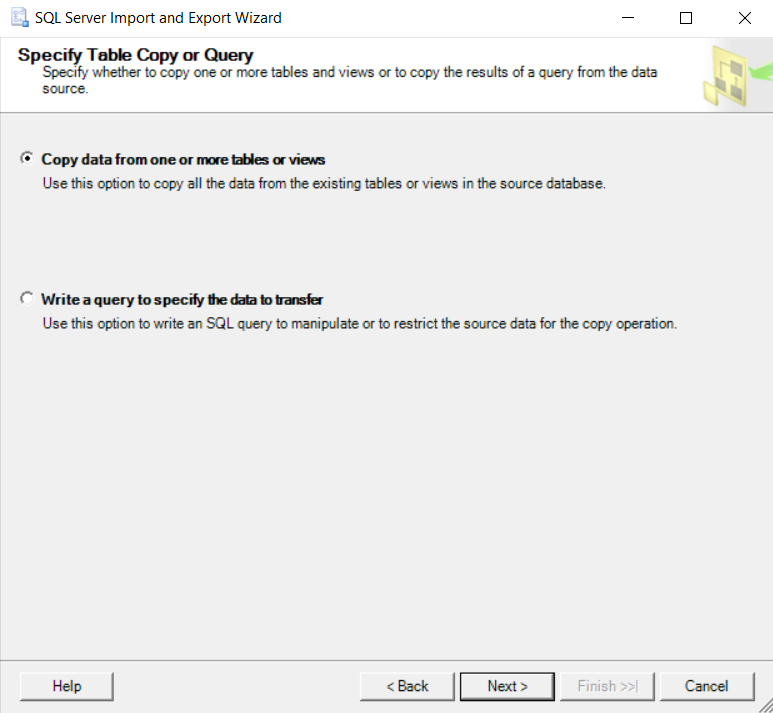
* Sau đó vào Start => SQL Server xxxx Import And Export Data (xxxx là phiên bản).
* Xuất hiện màn hình như sau, ta chọn file dữ liệu excel được lưu ở phần trước, và chọn phiên bản excel 2016 và next sang bước kế.



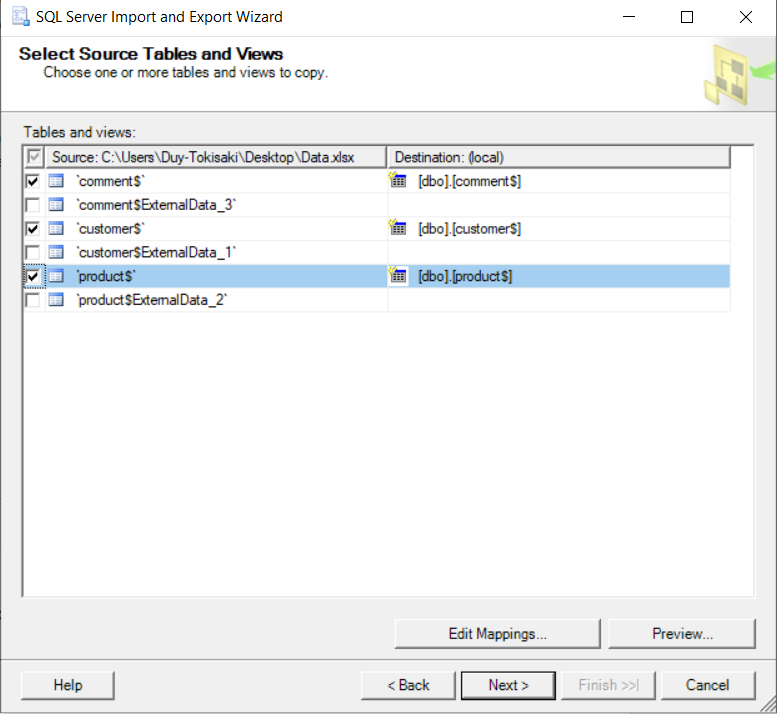
* Bước này ta chọn import vào Database có sẵn nên ta chọn SQL Server Native Client 11.0 và chọn Tiki vừa tạo ở bước trên.



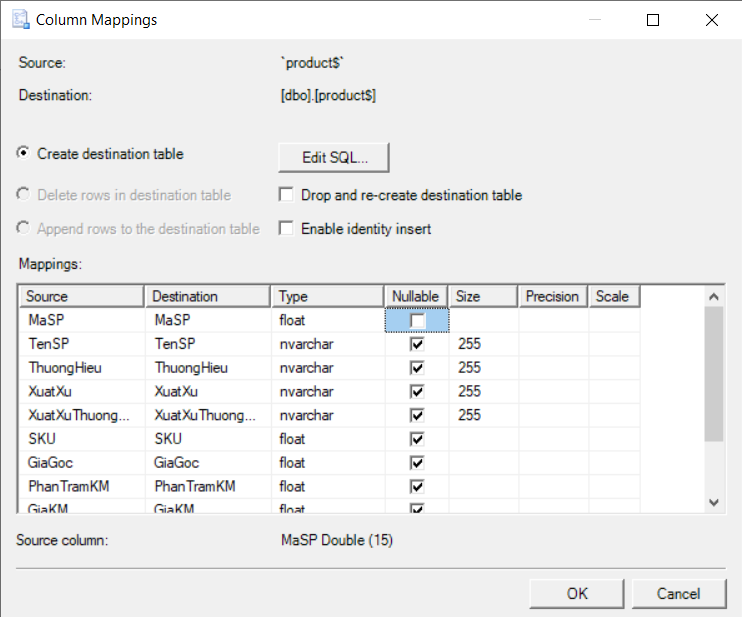
* Chọn Copy data from one or more tables or views.



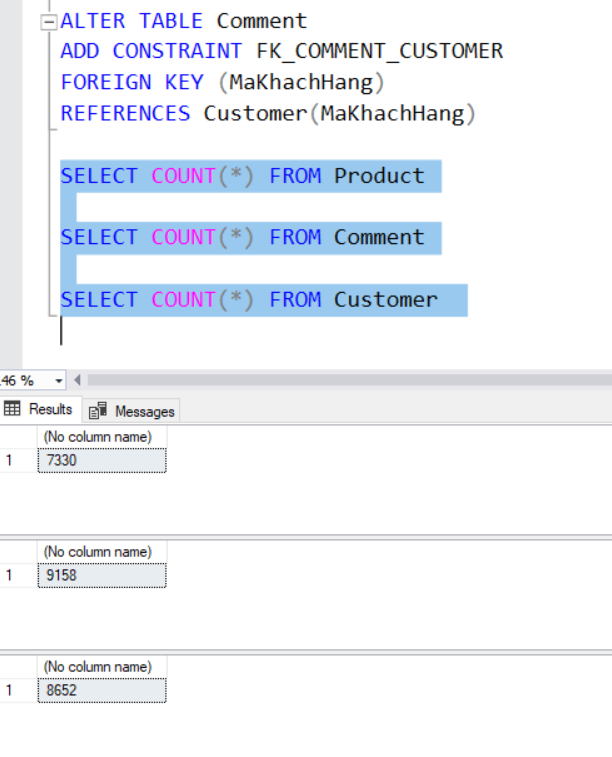
* Chọn các sheet cần import, đổi tên bảng cho phù hợp với yêu cầu, mặc định tên bảng là như hình.



* Chọn edit mappings => bỏ check của ô Nullable nếu cột đó là khóa chính.



* Kế đến các bạn chỉ cần next là được
* Tiếp theo cài đặt khóa chính và kiểm tra số dòng của các bảng.



1. **Hướng dẫn sử dụng source code:**

* Để mở source code, vào google colab 🡪 File 🡪 Upload NoteBook 🡪 chọn file .ipynb có trong thư mục nộp bài.

Ảnh có chứa văn bản

Mô tả được tạo tự động

* Để chạy code có 2 cách: chạy từng cell bằng tổ hợp phím Shift+Enter hoặc Runtime 🡪 Run All.

