**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**

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

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

**BẢNG THÔNG TIN CHI TIẾT**

|  |  |  |
| --- | --- | --- |
| **Mã nhóm:** | H2NP | |
| **MSSV** | **Họ tên** | **Email** |
| 1512361 | Bùi Thị Thanh Nguyệt | 1512361@student.hcmus.edu.vn |
| 1712462 | Trịnh Xuân Hội | 1712462@student.hcmus.edu.vn |
| 1712473 | Phạm Quang Hùng | 1712473@student.hcmus.edu.vn |
| 1712676 | Phạm Hồng Phúc | 1712676@student.hcmus.edu.vn |

**BẢNG PHÂN CÔNG**

|  |  |
| --- | --- |
| **Thành viên** | **Công việc** |
| 1512361 | Thực hiện tìm api, viết code, quay video |
| 1712462 | Thực hiện tìm api, viết code |
| 1712473 | Thực hiện tìm api, viết code |
| 1712676 | Thực hiện tìm api, viết code, viết báo cáo |

# **MÔ TẢ ĐỒ ÁN**

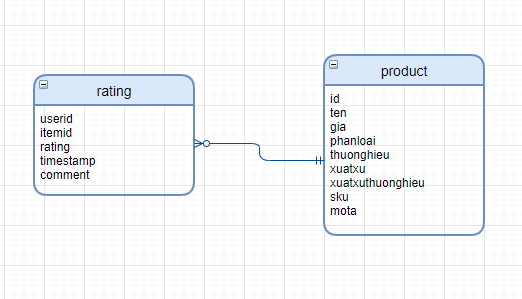
Thực hiện web crawler dữ liệu từ tiki chuẩn bị cho giai đoạn tư vấn theo chủ đề đã được phân công (Nhà cửa đời sống).

Thiết kế database gồm các thông tin về sản phẩm, rating và bình luận.

# **CƠ SỞ DỮ LIỆU**

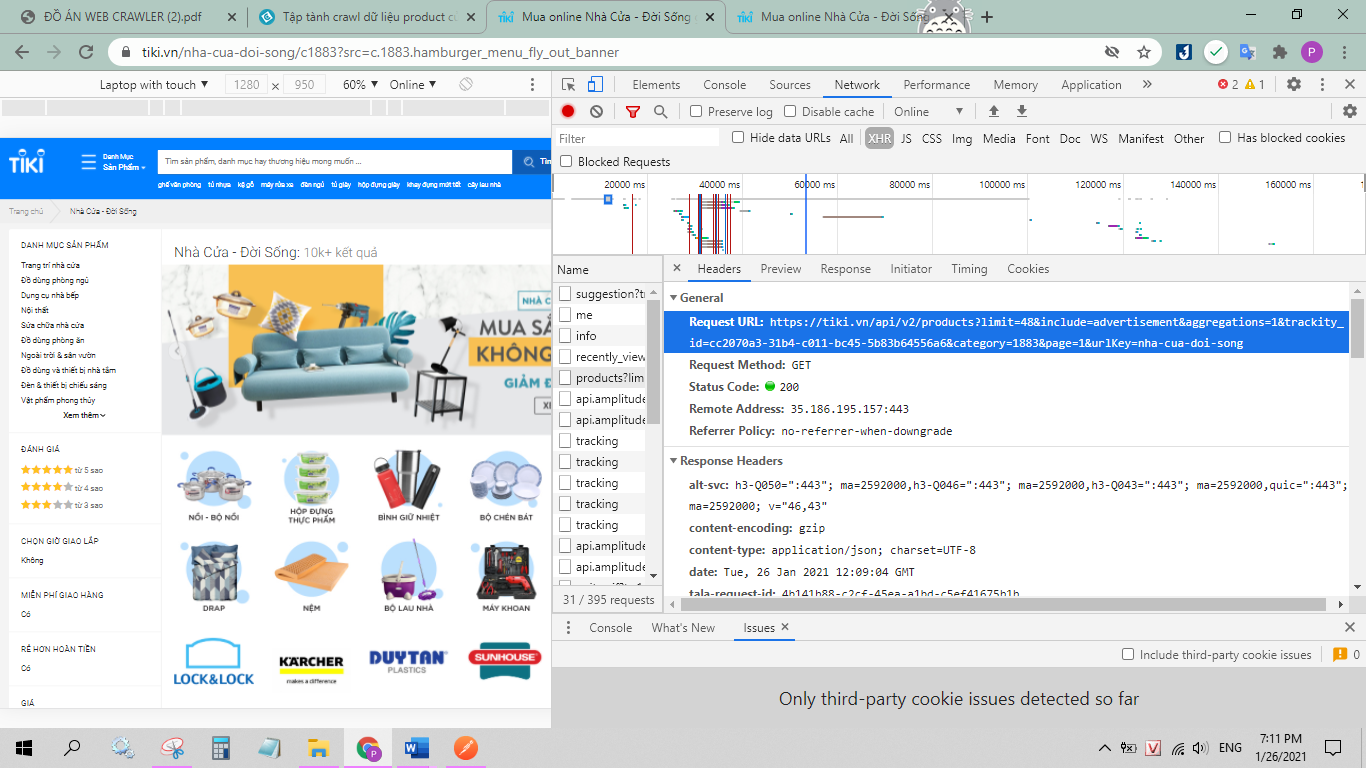
Gồm 2 bảng:

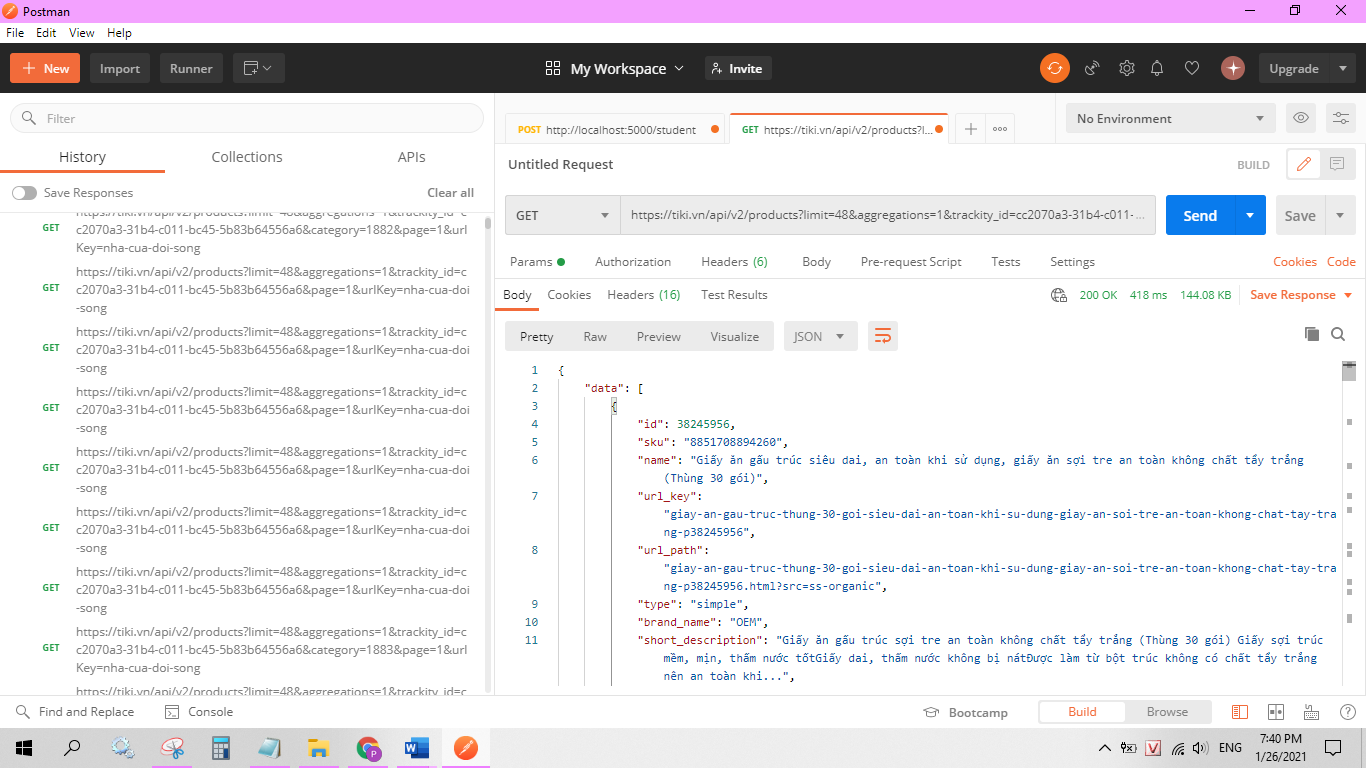
* Bảng “product” chứa thông tin sản phẩm (khóa chính là id).
* Bảng “rating” chứa thông tin đánh giá và bình luận của khách hàng cho sản phẩm (khóa chính là userid, itemid, timestamp).



# **CÁC BƯỚC THỰC HIỆN**

1. **Inspect Tiki và xác định thông tin cần lấy**

Sử dụng Chrome Developer Tool của Chrome. Chọn Network và chọn XHR để lọc các request AJAX, và xem xét hết tất cả các response để tìm xem cái nào sẽ trả những dữ liệu mà chúng ta cần.

Tìm được API danh sách các sản phẩm Nhà cửa đời sống <https://tiki.vn/api/v2/products?limit=48&aggregations=1&trackity_id=cc2070a3-31b4-c011-bc45-5b83b64556a6&category=1883&page=1&urlKey=nha-cua-doi-song>

Kiểm tra API vừa tìm được với postman

Tăng tham số page để lấy được tất cả sản phẩm (Có thể tăng limit để 1 lần lấy được nhiều sản phẩm)

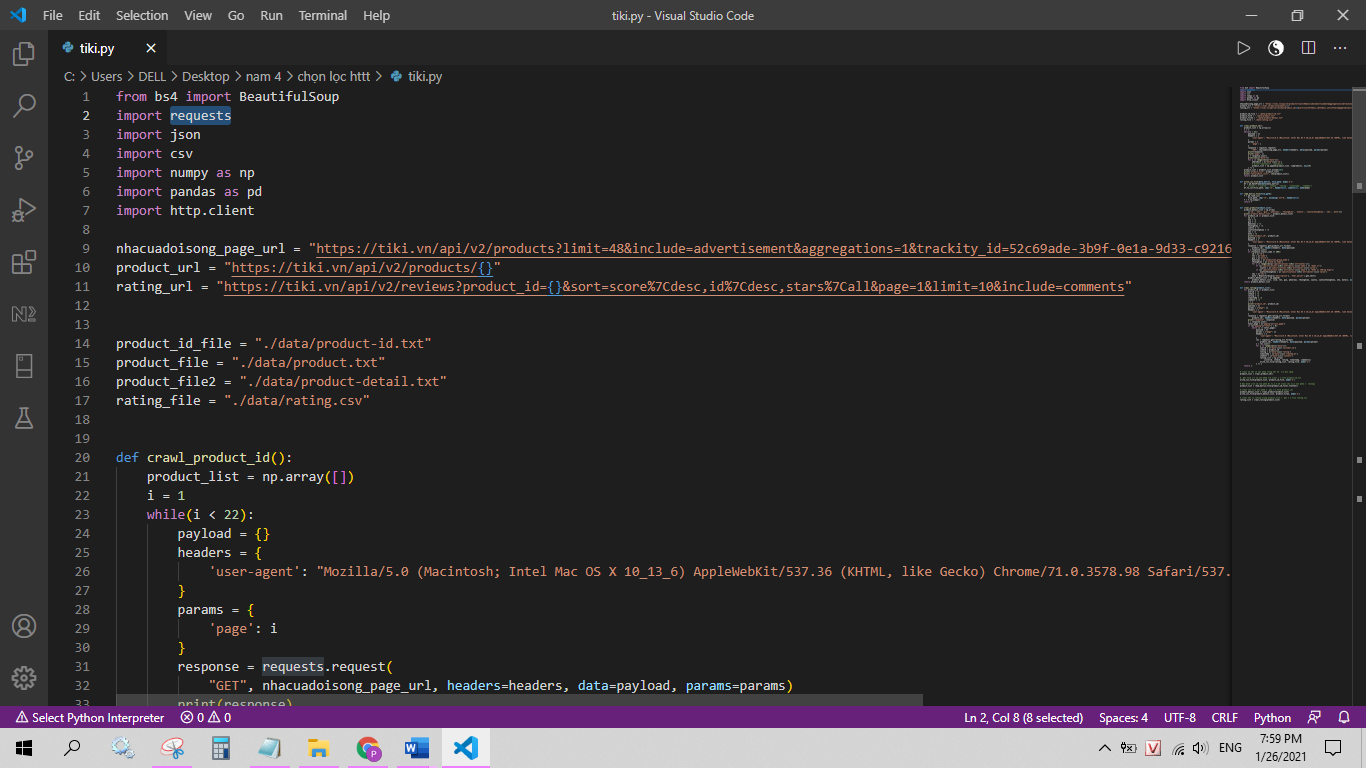
Vậy là đã xác định được một API dùng để lấy danh sách các sản phẩm.

Việc kế tiếp là tìm ra API :

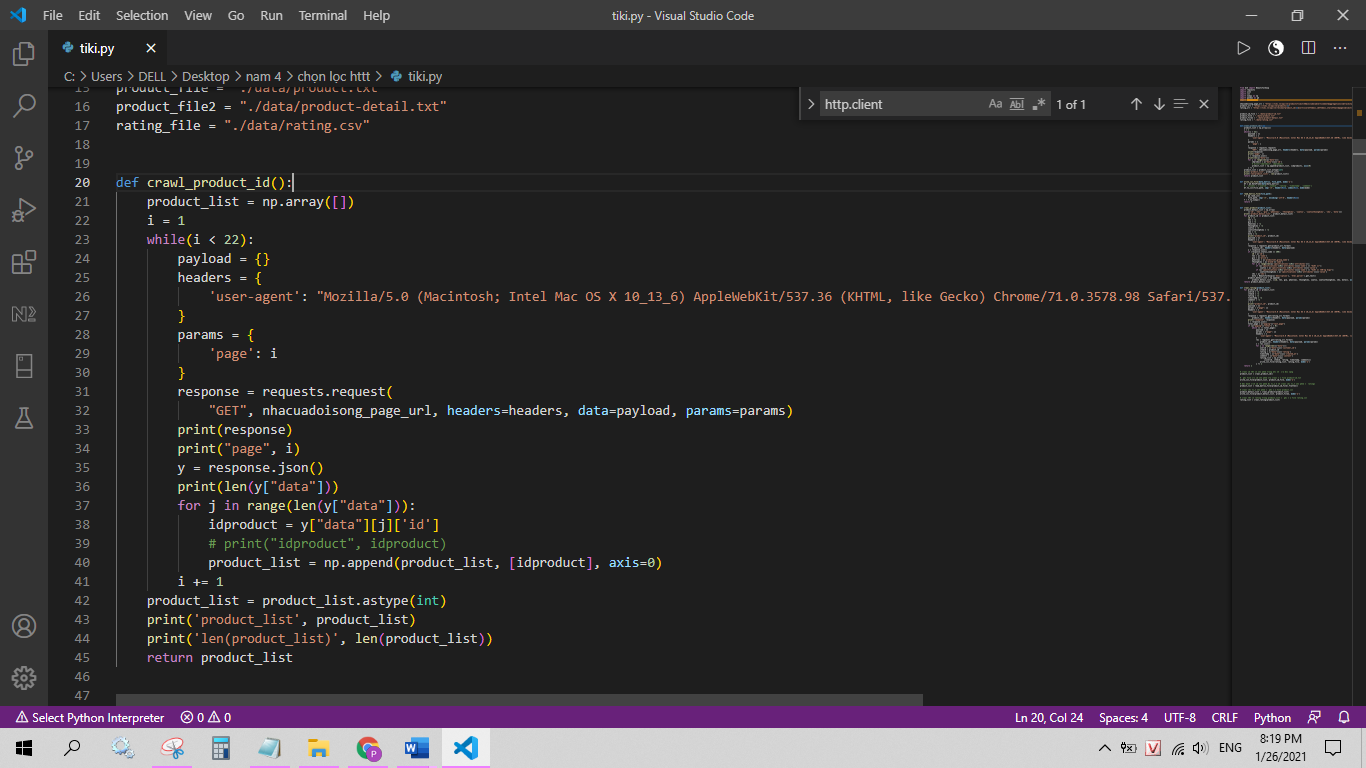
* Chứa thông tin của từng sản phẩm <https://tiki.vn/api/v2/products/38245956> (dùng id sản phẩm vừa lấy được từ danh sách trên).
* Chứa rating, bình luận và thời gian <https://tiki.vn/api/v2/reviews?product_id=38245956&sort=score%7Cdesc,id%7Cdesc,stars%7Call&page=1&limit=10&include=comments>

1. **Crawl dữ liệu trên Tiki**

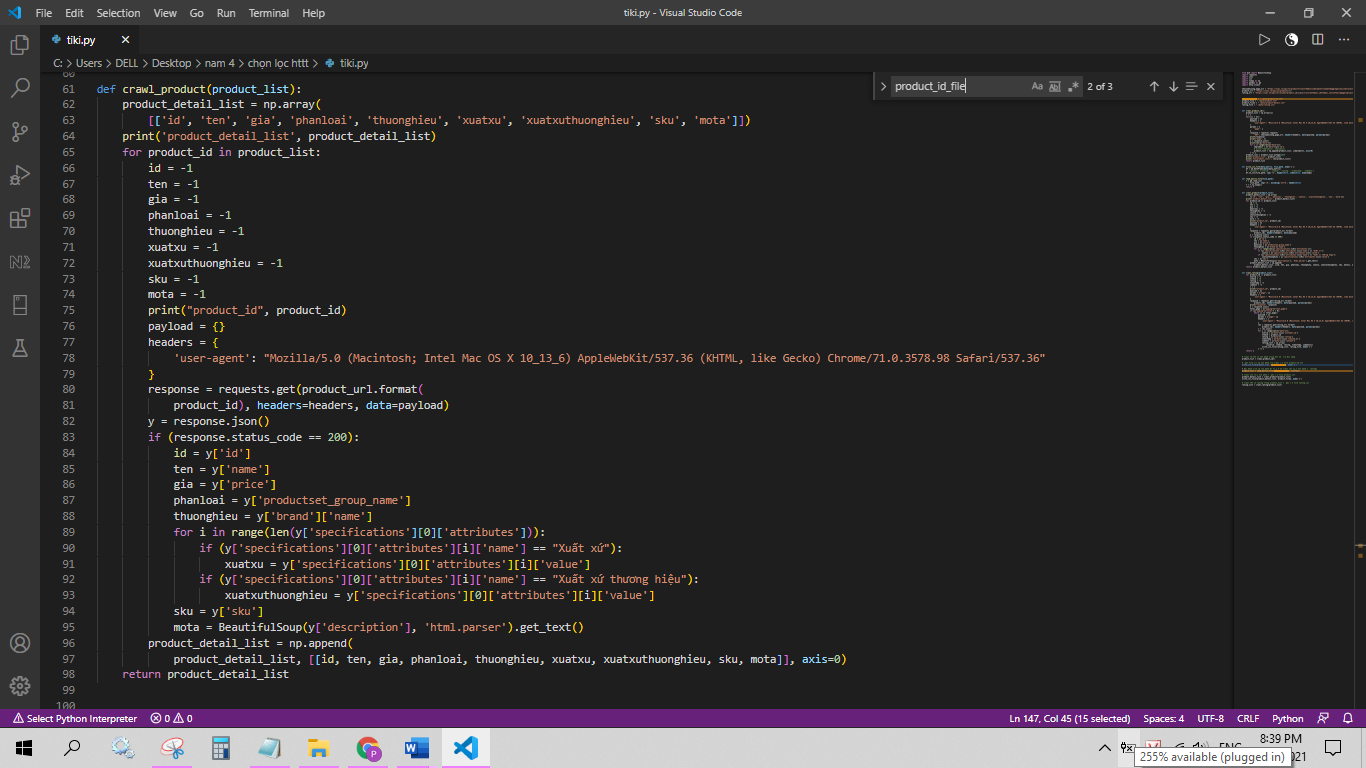
**Các thư viện sử dụng**

* Thư viện BeautifulSoup trên Python để đọc các tag HTML lấy dữ liệu.
* Thư viện Requests trong Python giúp lập trình viên có thể thực hiện các tác vụ như gửi request tới server cũng như xử lý response một cách đơn giản.
* Thư viện json cung cấp những công cụ để làm việc cũng như để mã hóa và giải mã dữ liệu JSON.
* Module csv để xử lý đọc ghi file csv.
* Thư viện numpy để xử lý các mảng.
* Thư viện pandas để xử lý dữ liệu trên cấu trúc dataframe.

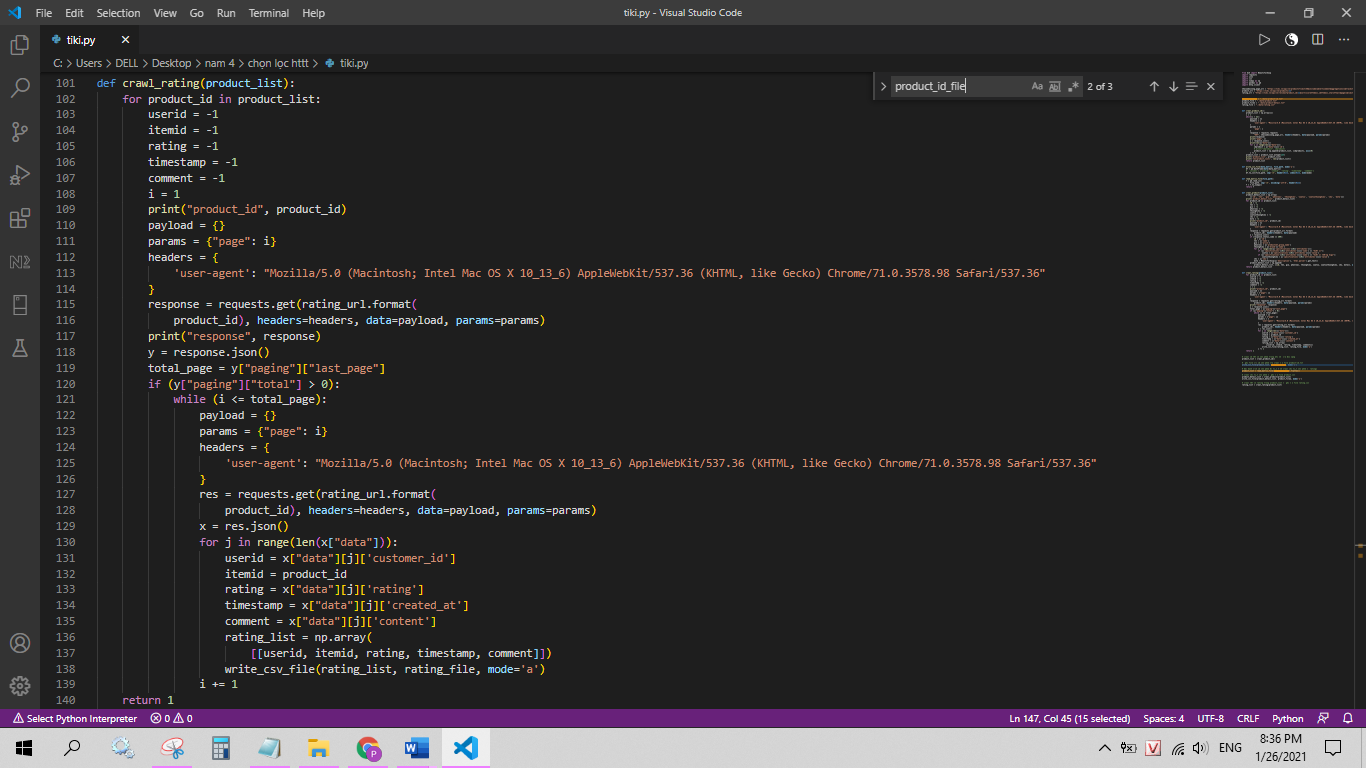
**Crawl hết tất cả mã sản phẩm**



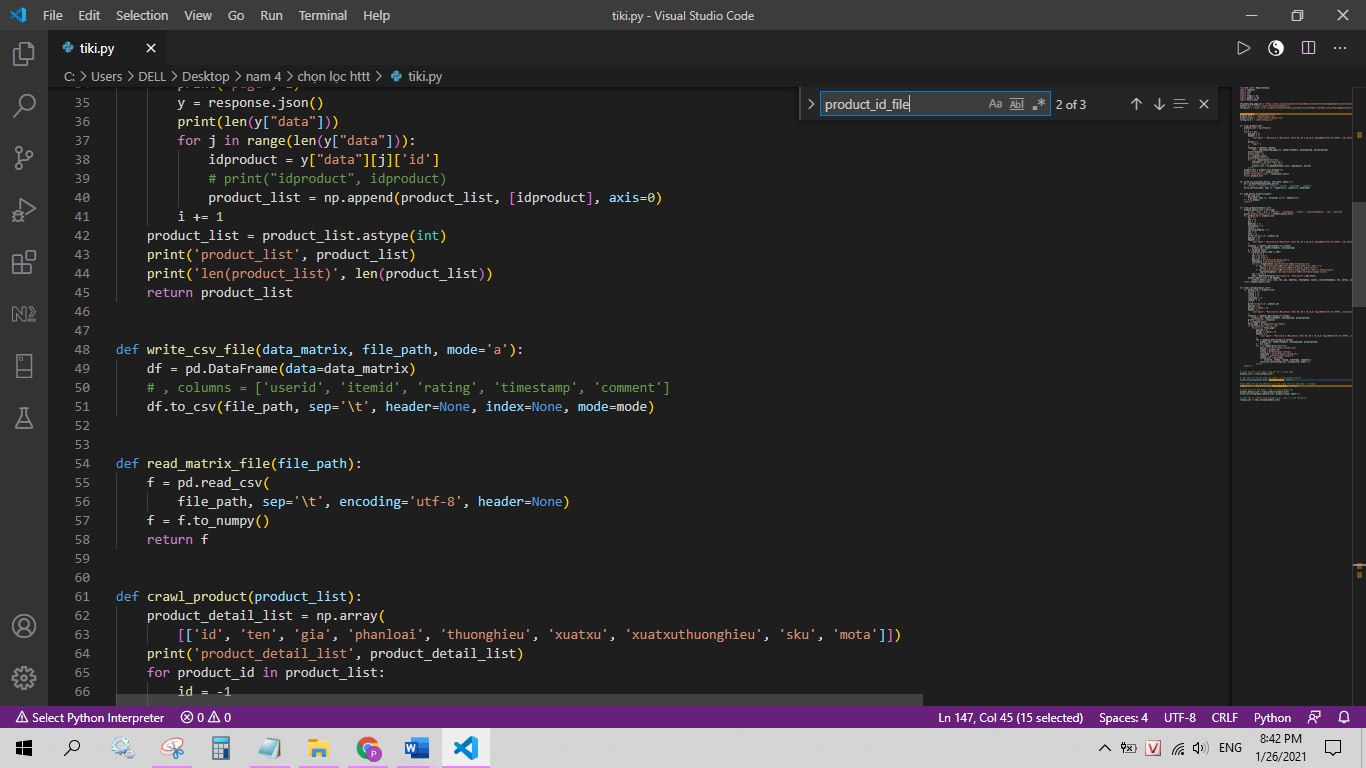
**Crawl tất cả chi tiết sản phẩm**



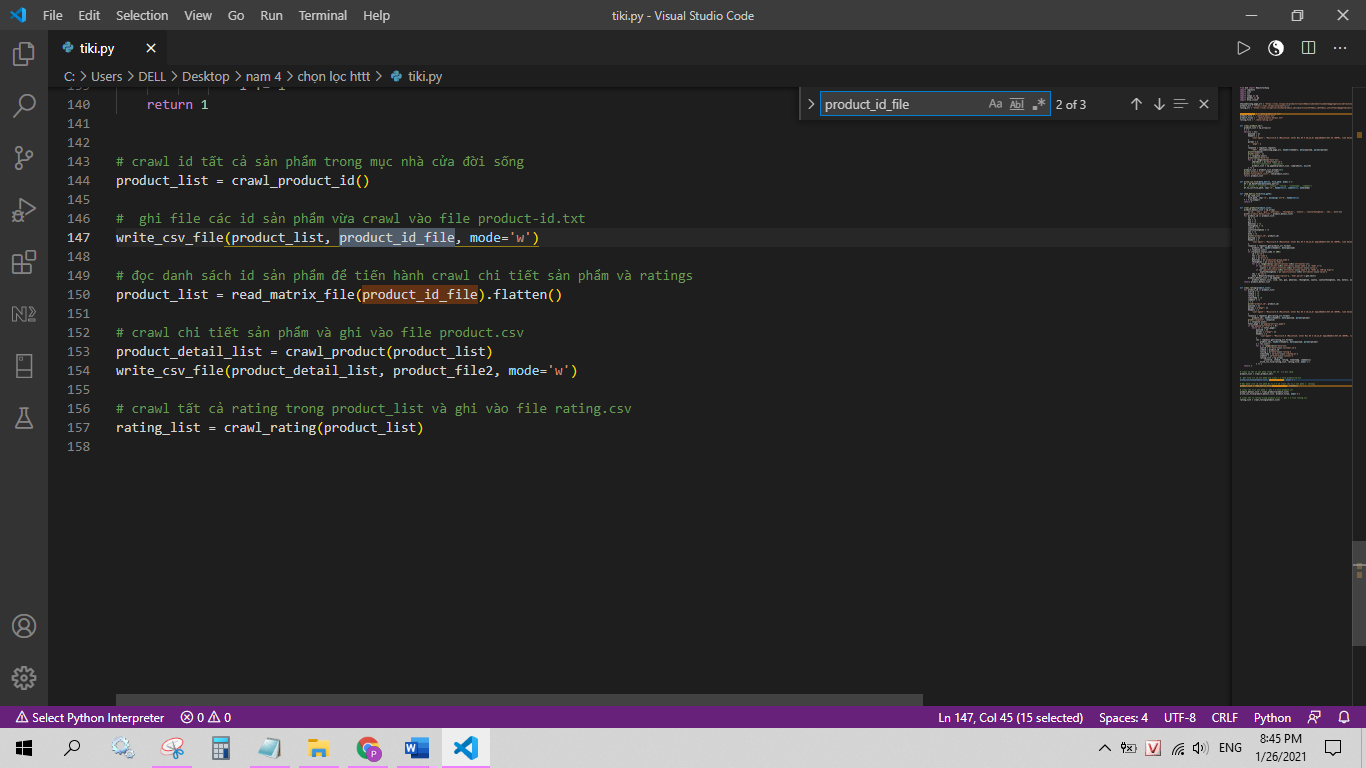
**Crawl các rating và bình luận của sản phẩm**



**Hàm đọc file và ghi dữ liệu vừa crawl được**

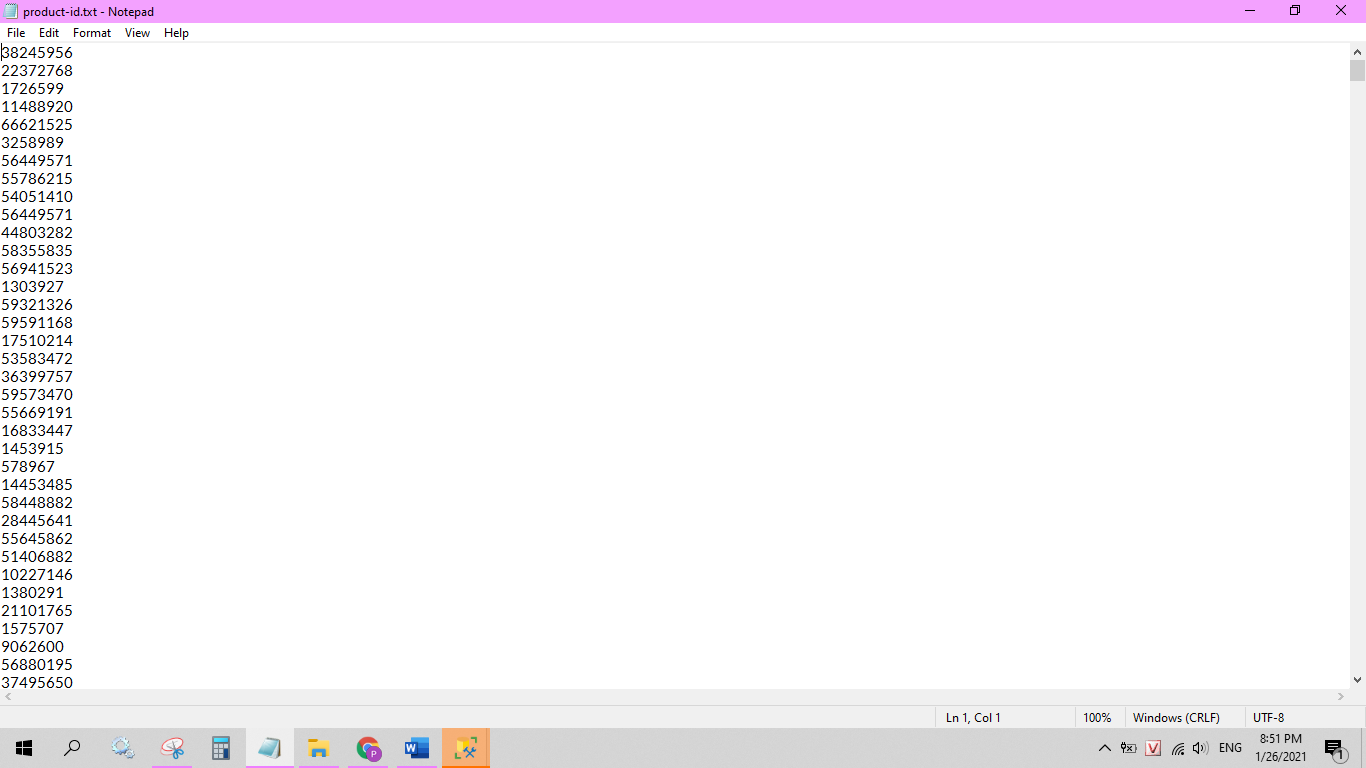


**Thực hiện crawl dữ liệu**

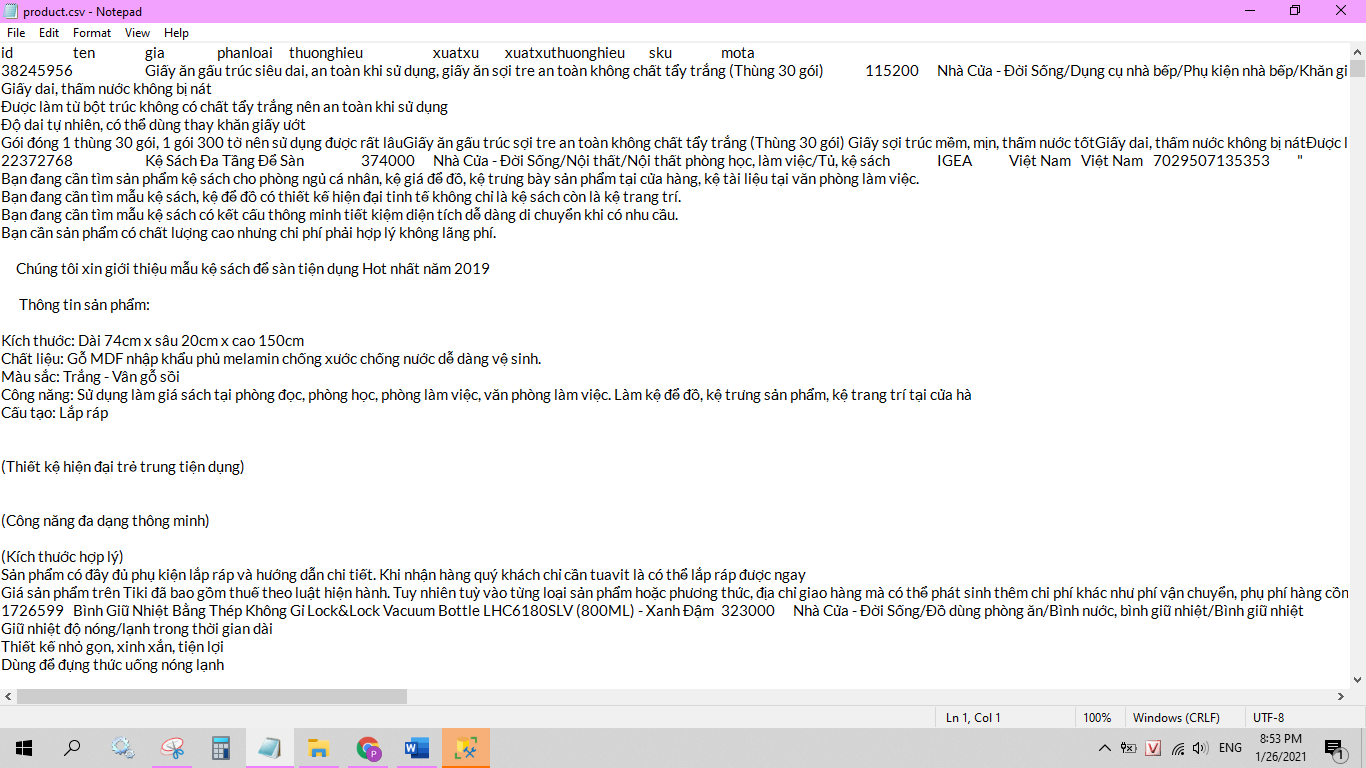


**Kết quả crawl dữ liệu**

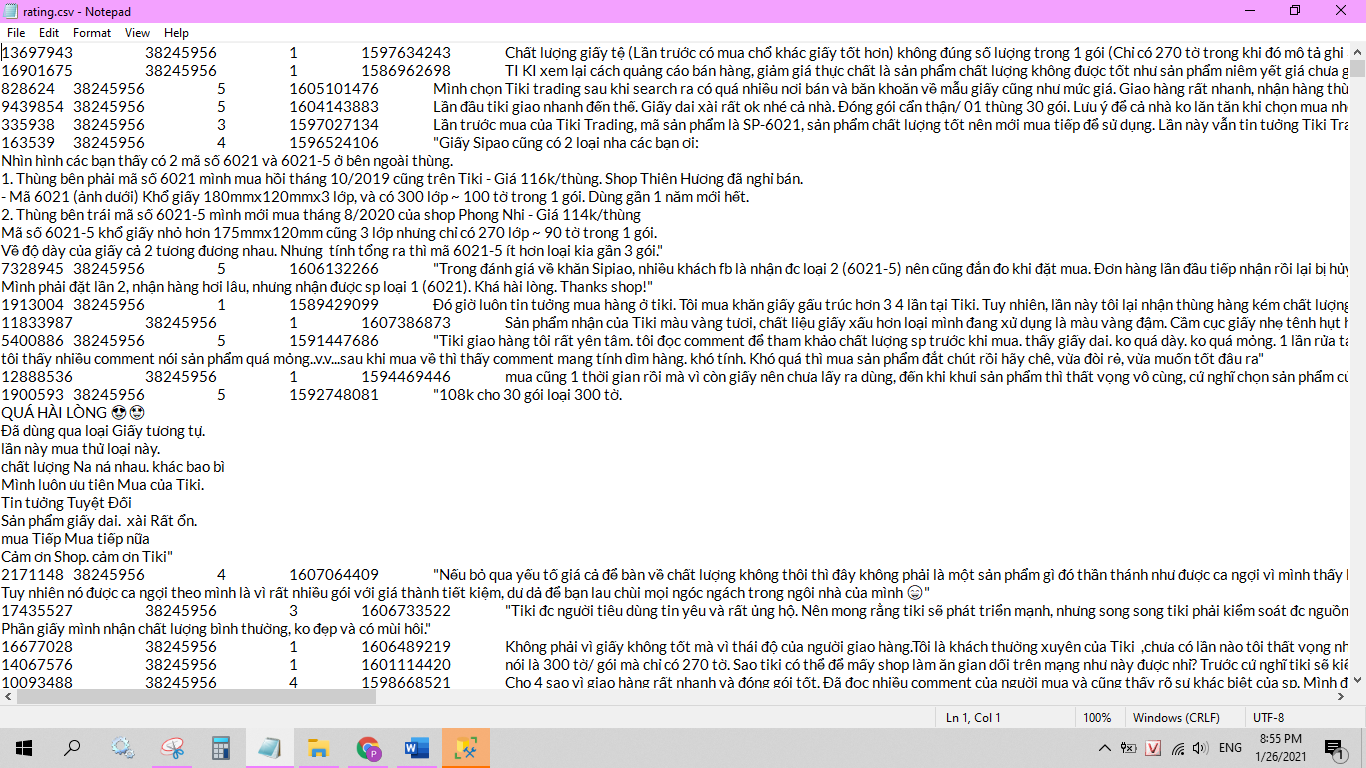
Danh sách id sản phẩm



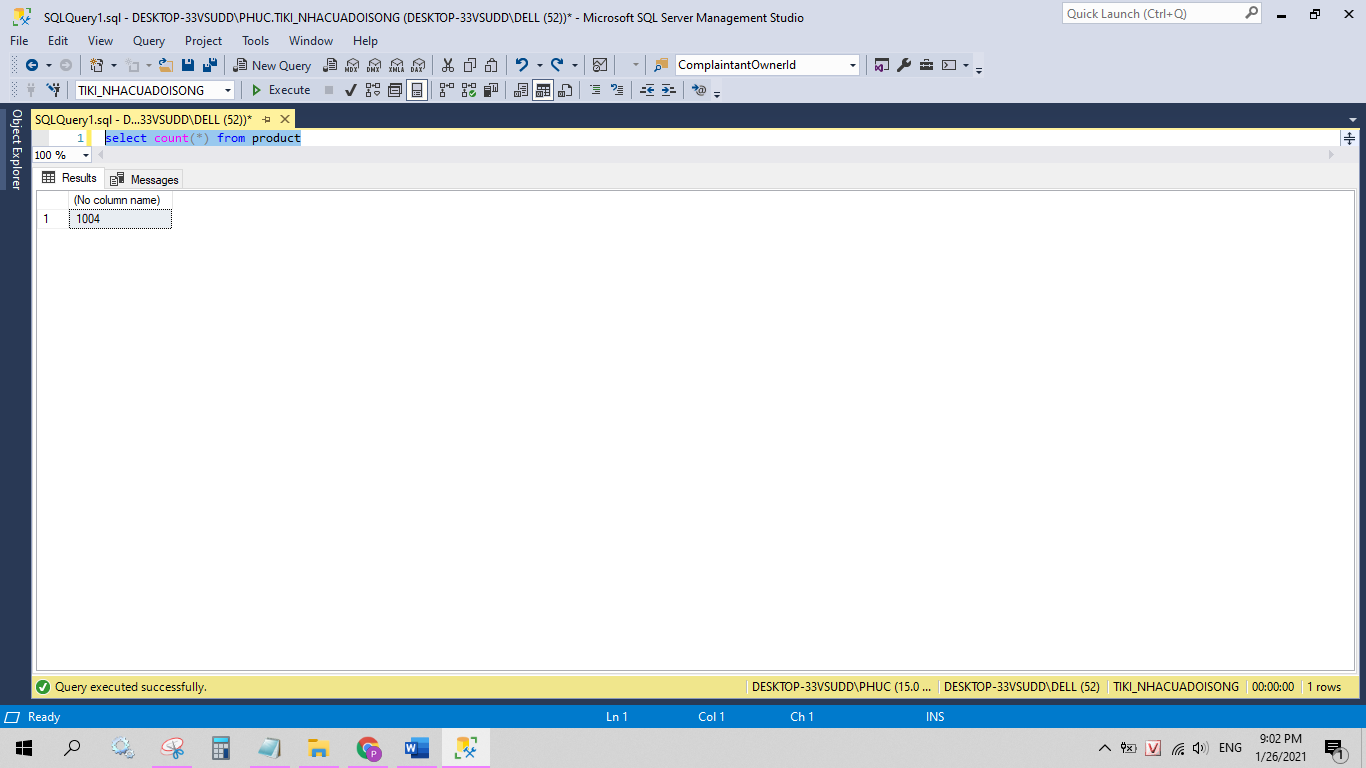
Danh sách chi tiết sản phẩm

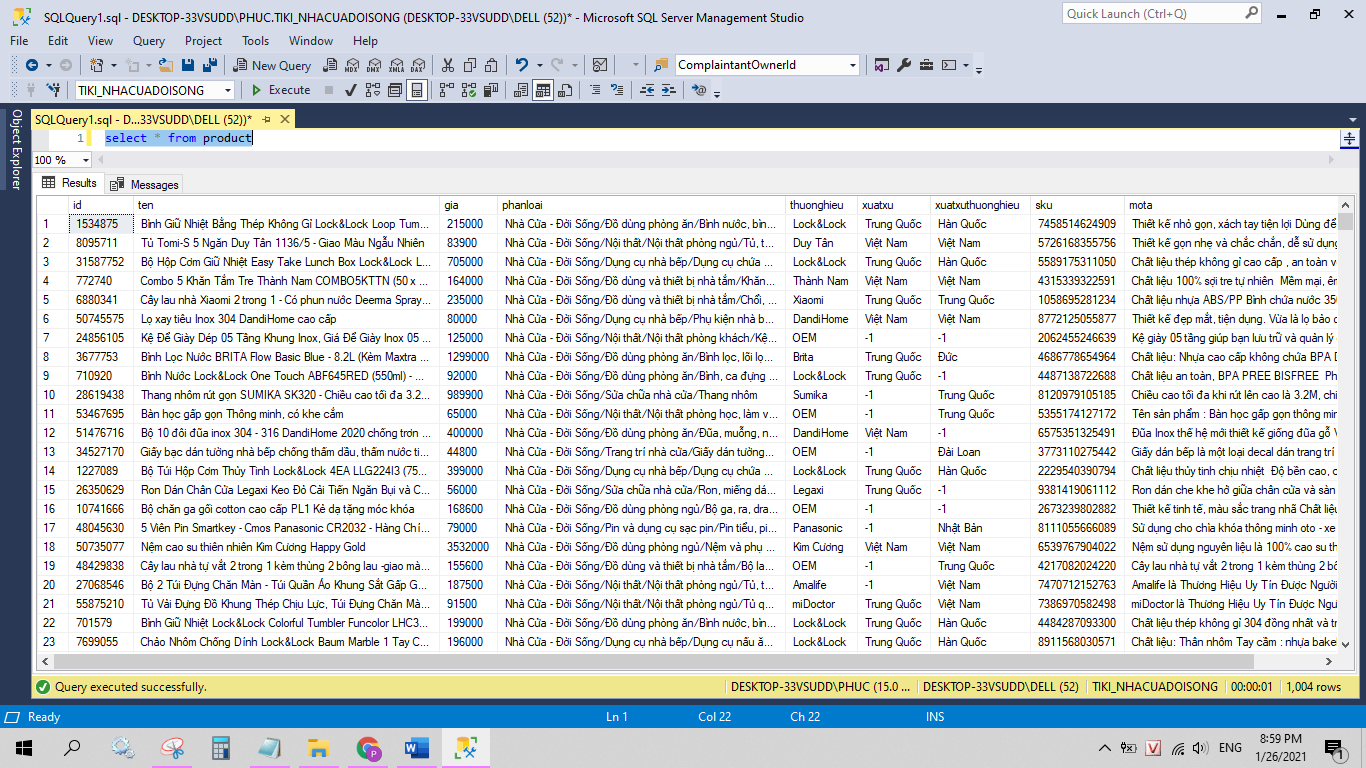


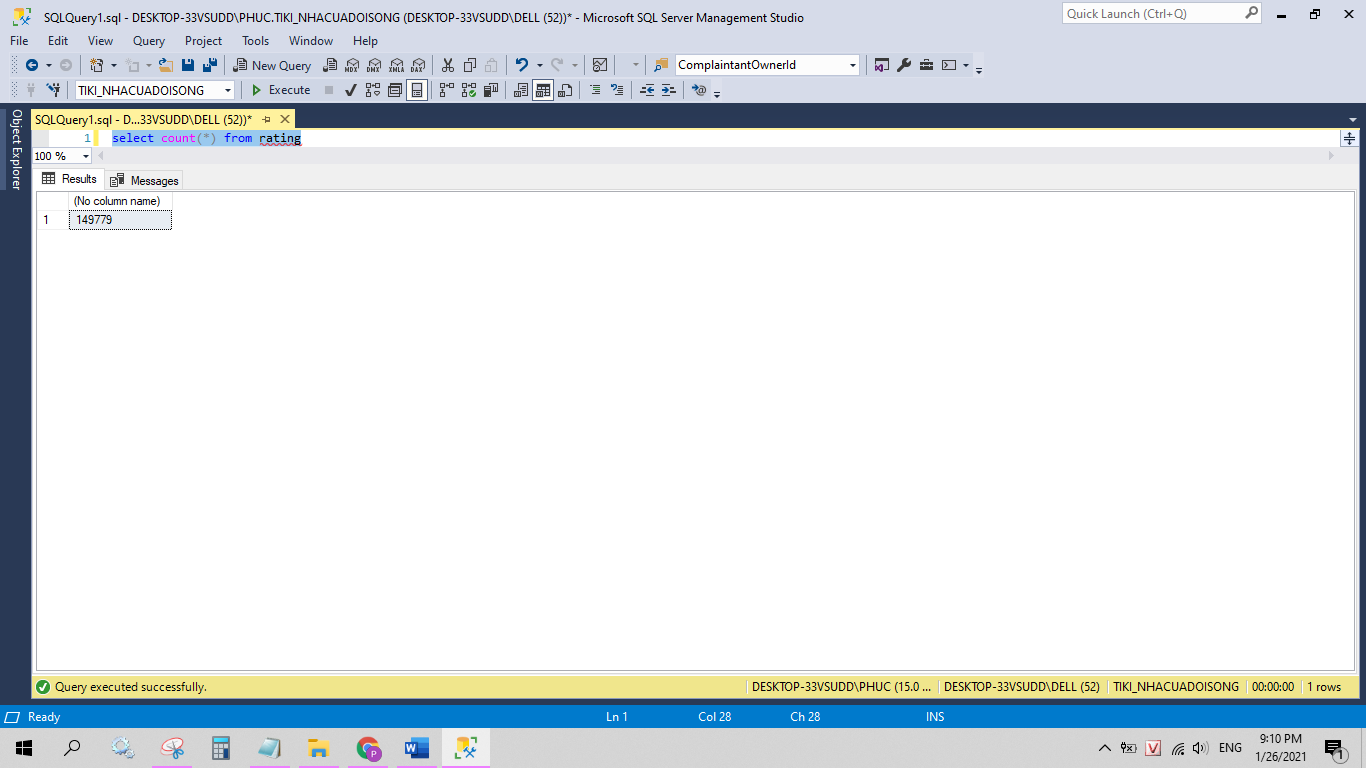
Danh sách rating và bình luận



**Kết quả import cái file trên vào CSDL**

Bảng product  crawl được tất cả **1004** sản phẩm Nhà ở đời sống



Bảng rating  crawl được **149 779** dòng rating–bình luận các sản phẩm Nhà cửa đời sống

