A blue and white logo

Description automatically generated

Trường Đại học Khoa học Tự nhiên – ĐHQG HCM

Ngành: Công nghệ Thông tin – Chất lượng cao

Lớp: 20KHMT

Môn: Khoa học Dữ liệu Ứng dụng

Thành viên tham gia:

1. Hoàng Phước Gia Nguyên 20127574
2. Nguyễn Thiện Nhân 20127265
3. Nguyễn Mạnh Cường 20127456

Giảng viên hướng dẫn:

1. Thầy Lê Ngọc Thành
2. Thầy Nguyễn Tiến Huy
3. Cô Nguyễn Thị Thu Hằng

**PROGRESS REPORT - MIDTERM**

# Tổng quan đồ án

Giới thiệu đồ án: Xây dựng hệ thống gợi ý sản phẩm cho người dùng từ tập dữ liệu sản phẩm của nền tảng thương mại điện tử [Tiki](https://tiki.vn).

Đầu tiên, tập dữ liệu sản phẩm theo các category sau sẽ được crawl từ trang chủ Tiki thông qua APIs: “[Thời trang nam](https://tiki.vn/thoi-trang-nam/c915)”, “[Thời trang nữ](https://tiki.vn/thoi-trang-nu/c931)”, “[Giày dép nam](https://tiki.vn/giay-dep-nam/c1686)”, “[Giày dép nữ](https://tiki.vn/giay-dep-nu/c1703)” và lưu vào file CSV (Comma-Separated Values).

Mục tiêu của đồ án: Hoàn thiện một mô hình học máy giúp đề xuất các sản phẩm thời trang phù hợp theo sở thích của người dùng cuối.

# Báo cáo tiến độ và dự kiến

* 1. Crawl dữ liệu
     1. *Tìm danh mục và API*

Truy cập vào trang chủ Tiki.vn, bật Inspect (click chuột phải chọn Inspect) và chuyển sang mục Network ở thanh Taskbar trên cùng.

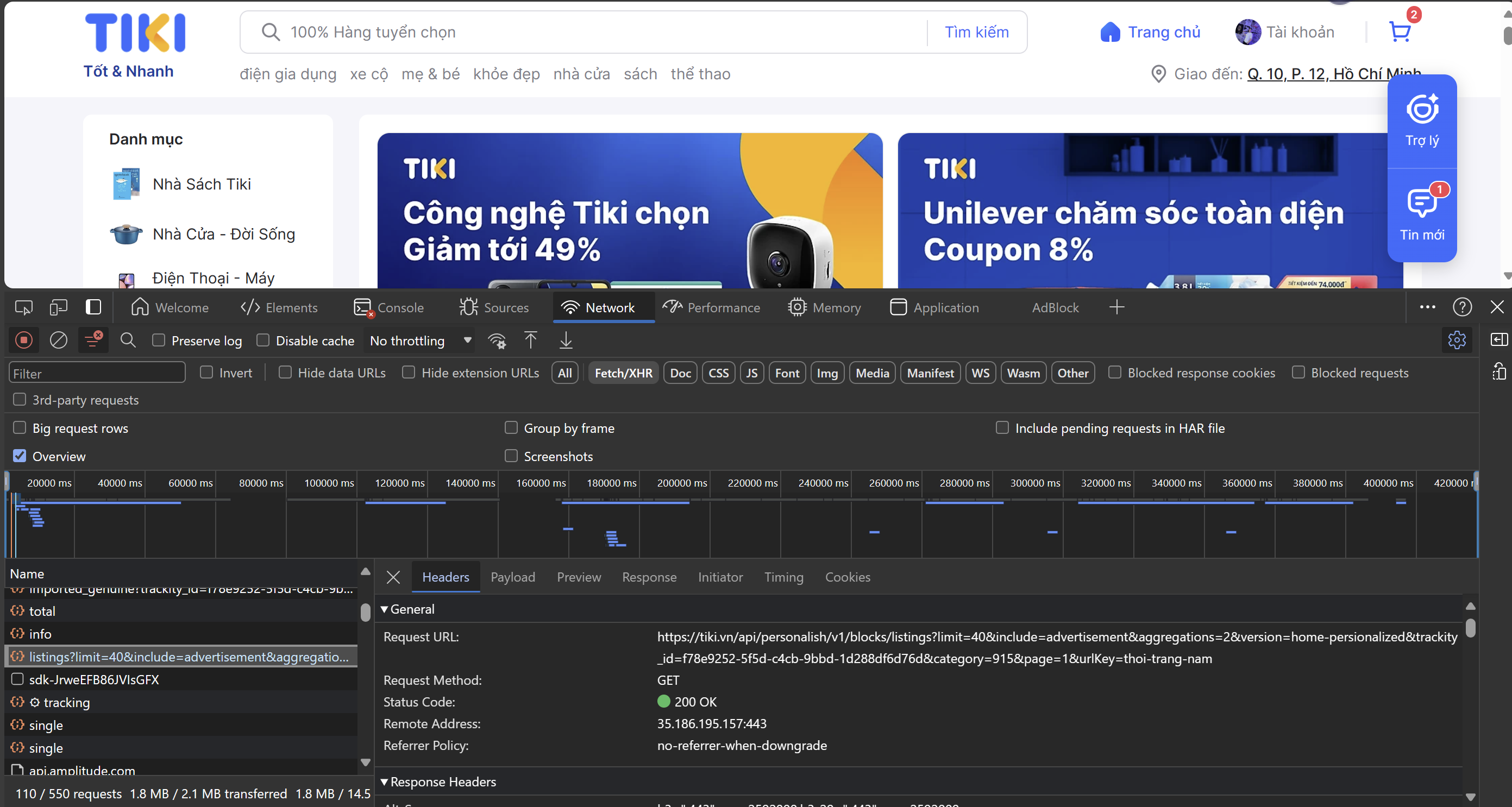


Figure 1: Hình ảnh trang Inspect ở trang chủ

Sau đó, nhấp chọn vào phần Danh mục, chọn lần lượt 1 trong 4 danh mục tương ứng đã được liệt kê ở mục (1).

A screenshot of a computer

Description automatically generated

Figure 2: Lựa chọn 1 danh mục thời trang. Ở đây đầu tiên lựa chọn 'Thời trang nam'

Lọc dữ liệu theo Fetch/XHR và tìm kiếm mục ‘listings?limit=40…’. Đây là đường API để lấy dữ liệu của danh mục hiện tại.

A screenshot of a computer

Description automatically generated

Figure 3: Tìm API của Tiki để lấy dữ liệu sản phẩm trong danh mục

Double click chuột trái vào đường API (hoặc sử dụng ứng dụng Postman) để xem dữ liệu các sản phẩm có trong danh mục đang cần crawl.

A black background with orange and black lines

Description automatically generated

Figure 4: Minh họa thông tin của một sản phẩm trong danh mục

* + 1. *Crawl id của sản phẩm*

Lấy các header, cookies và params cần thiết để tiến hành việc crawl

A screenshot of a computer program

Description automatically generated

Figure 5: Chuẩn bị các tham số cần thiết cho việc crawl dữ liệu

Tiến hành việc crawl các id của tất cả sản phẩm trong danh mục hiện tại và lưu vào file ‘product\_id\_ncds.csv’

A screenshot of a computer

Description automatically generated

Figure 6: Các id của một danh mục lưu vào product\_id\_ncds.csv

* + 1. *Crawl sản phẩm*

Sau khi có được danh sách id sản phẩm, thực hiện việc crawl các thông tin cần thiết (id, sku, name, short\_description, price, list\_price, discount, discount\_rate, review\_count, rating\_average, inventory\_status, quantity\_sold, bran\_id, brand\_name, spid, seller\_id) của sản phẩm và lưu vào file csv có tên <tên-danh-mục>.csv

A screenshot of a computer

Description automatically generated

Figure 7: Minh họa một bảng thông tin các sản phẩm thuộc danh mục 'Giày dép nam' (mã hóa UTF-8)

Các file đã crawl thông tin sản phẩm bao gồm:

* Thoi\_trang\_nam.csv
* Thoi\_trang\_nu.csv
* Giay\_dep\_nam.csv
* Giay\_dep\_nu.csv

Ngoài ra, trong quá trình crawl, đã xảy ra lỗi không thể đọc một số dữ liệu JSON của một sản phẩm trong danh mục. Các id của các sản phẩm này đã được lưu lại trong các file csv có tên <tên-danh-mục\_error\_product\_list.csv>.

A screenshot of a computer

Description automatically generated

Figure 8: Minh họa danh sách các sản phẩm hiện bị lỗi định dạng JSON chưa crawl được

Trong đó, có 4 file có tên <tên-danh-mục\_error\_product\_list.csv> tương ứng với 4 danh mục đã được nêu ở phần (1).

* + 1. *Crawl số sao đánh giá*

Sau khi có được thông tin các sản phẩm trong danh mục, tiếp tục crawl số sao đánh giá của từng sản phẩm trong <tên-danh-mục.csv> dựa theo product\_id, spid và seller\_id. Các sản phẩm này phải là các sản phẩm không nằm trong <tên-danh-mục\_error\_product\_list.csv> tức không xảy ra lỗi khi crawl.

A screenshot of a computer

Description automatically generated

Figure 9: Minh họa số sao đánh giá của các sản phẩm (1-5 sao)

Hàm xử lý sẽ đếm số lượng 1-5 sao (kèm phần trăm) của từng sản phẩm cùng với đánh giá trung bình và lưu vào các file <tên-danh-mục\_stars.csv>

* 1. **Tiền xử lý dữ liệu (Preprocessing)**

Sử dụng các hàm parser để bắt các lỗi xuất hiện trong quá trình crawl dữ liệu sản phẩm <tên-danh-mục.csv> và dữ liệu đánh giá <tên-danh-mục\_stars.csv>

A screen shot of a computer program

Description automatically generated

Sau quá trình tiền xử lý dữ liệu, ghi nhận được

* Thời trang nam: 1390 sản phẩm và 1390 số sao tổng quát tương ứng, 640 id lỗi
* Thời trang nữ: 1267 sản phẩm và 1267 số sao tổng quát tương ứng, 763 id lỗi
* Giày dép nam: 1240 sản phẩm và 1240 số sao tổng quát tương ứng, 790 id lỗi
* Giày dép nữ: 1173 sản phẩm và 1173 số sao tổng quát tương ứng, 857 id lỗi
  1. **Xây dựng mô hình học máy**
  2. **Đánh giá mô hình**

# Phân công công việc

|  |  |  |
| --- | --- | --- |
| Công việc | Thành viên thực hiện | Tiến độ thực hiện |
| Crawl dữ liệu sản phẩm | Hoàng Phước Gia Nguyên + Nguyễn Thiện Nhân | Đã hoàn thành |
| Crawl số sao sản phẩm | Nguyễn Mạnh Cường | Đã hoàn thành |
| Tiền xử lý dữ liệu | Cả team | Đã hoàn thành |
| Xây dựng mô hình học máy | Hoàng Phước Gia Nguyên + Nguyễn Thiện Nhân | Chưa hoàn thành |
| Đánh giá mô hình | Nguyễn Mạnh Cường | Chưa hoàn thành |