

**Đồ án cuối kỳ: Phân tích thị trường E-Food**

TỔNG HỢP CÁC NOTEBOOK

| Giáo viên hướng dẫn:  Nguyễn Tiến Huy | Nhóm K18:  18120203 - Vũ Đăng Hoàng Long  18120212 - Huỳnh Long Nam  18120249 - Hồ Minh Trí  18120462 - Nguyễn Thị Mận  18120505 - Đào Quốc Phong  18120544 - Trương Minh Tài  18120579 - Đặng Minh Thọ |
| --- | --- |

# Foody

**Cào URL & Cào thông tin quán ăn:**

*1.1 - Foody - hcmc\_foody.ipynb*

Input: URL '<https://www.foody.vn/ho-chi-minh/dia-diem>' với các tham số xác định được lấy từ Inspect → Network

Output: File thông tin của các quán ăn trên trang web Foody

(**hcmc-0-101000-update-3.csv**)

**Cào bình luận quán ăn:**

*1.2 Foody - Crawl cmt.ipynb*

Input: Danh sách restaurant\_id (VD **url\_40k.csv**)

Output: File thông tin chi tiết bình luận (**cmt\_data\_full.csv**)

# Shopee Food

**Quét các tọa độ của TP HCM:**

*2.1 Scan location.ipynb*

Output: 2 file **ids.xlsx** và **location.xlsx**

**Cào thông tin quán ăn:**

*2.2 Shopee - Crawl restaurant.ipynb*

Input: **ids.xlsx** và **location.xlsx**

Output: thông tin chi tiết của quán ăn **out\_split\_17.xlsx**

**Cào thông tin món ăn:**

*2.3 Shopee - Crawl dishes.ipynb*

Input: danh sách restaurant\_id theo quận

**raw data/retaurant\_id/{quận}.xlsx**

Output: thông tin chi tiết các món ăn **raw data/dish/{quận}\_dishes.xlsx**

**Cào các món ăn kèm của từng món ăn:**

*2.4 Shopee - CrawlToppingData.ipynb*

Input: **DIM\_SHOPEE\_DISHES** từ **Main.db**

Output: File thông tin các món ăn kèm của từng món ăn (**topping\_data1.csv**).

**Tối ưu cào thông tin quán ăn và món ăn:**

*2.5 Shopee - Optimize crawl restaurant and dish.ipynb*

Input: **Main.db**

Output: các file parquet thô gồm nhiều schema trong **raw data/shopee dish parquet/{date}\_dishes**

(VD **20221129\_dishes**)

**Khám phá và đồng nhất các schema parquet:**

*2.6 Shopee - EDA and Unify parquet.ipynb*

Input các file kết quả của bước 2.6

Output: các file parquet thô đã được hợp nhất schema trong **cleaned data/Staging/{date}.dishes.parquet** (VD **2022.11.26.dishes.parquet**)

**Khám phá và trích xuất các cột cần thiết từ parquet:**

*2.7 Shopee - EDA and Extract parquet.ipynb*

Input các file kết quả của bước 2.7

Output: các file parquet đã được chọn lọc cột và ở dạng bảng trong **cleaned data/Foundation** dạng **{date}.restaurant\_dishes.parquet** hoặc **{date}.restaurant\_catalogs.parquet** chứa thông tin tương ứng về món ăn hoặc catalog

VD:

* **2022.12.03.restaurant\_catalogs.parquet**
* **2022.12.03.restaurant\_dishes.parquet**

# Tổng hợp dữ liệu:

**Hợp nhất dữ liệu Shopee, Foody và mô hình hóa dữ liệu**

*3.1 Data Modeling.ipynb*

Input:

Thông tin chi tiết các quán Foody (**hcmc-0-101000-update-3.csv**)

Thông tin chi tiết các quán Shopee (**hcmc\_shopee\_order\_like.xlsx**)

Thông tin chi tiết các món ăn Shopee (**raw data/dish/{quận}\_dishes.xlsx**)

Output: **Main.db** gồm các bảng: Active hours, Audience, Category, Brand, Cuisne, Dishes, Root Category, Restaurant

Do dữ liệu có nhiều rác, cần phải sửa thủ công (sửa thương hiệu, sửa tọa độ, sửa giá món ăn bất thường) bằng nhiều phương pháp, nên không thể include tất cả các bước trên trong notebook

# Thử nghiệm mô hình

**TF-IDF**

*4.1 insight trà sữa*

Input: **Main.db**

Output: các quán là quán trà sữa (**restaurant\_groupby.csv)**

**Transform dữ liệu và thử nghiệm mô hình**

*4.2 Transform and Model Experiment.ipynb*

Input:

* **Main.db**
* File dishes.parquet của ngày muốn so sánh (VD **2022.11.26.restaurant\_dishes.parquet**)
* Danh sách các quán trà sữa (**restaurant\_groupby.csv**)

Output: Model và các kết quả thử nghiệm