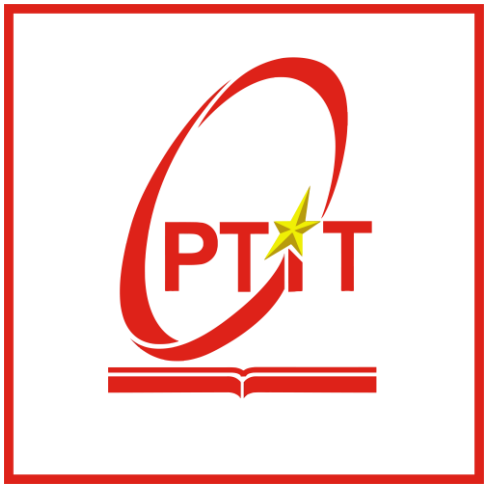
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Báo cáo hàng tuần

Môn học: Thực tập cơ sở

Giảng viên: Kim Ngọc Bách

Họ tên: Trần Quang Anh

Mã SV: B22DCCN044

Lớp: E22CQCN04-B

Tuần: 2

Báo cáo thực tập cơ sở - Tuần 2: Nguồn Dữ Liệu và Tiền Xử Lý Dữ liệu cho Hệ Thống Gợi Ý Món Ăn

I. Dữ Liệu Mong Muốn cho Hệ Thống

1. Mục Tiêu của Dữ Liệu

Để xây dựng một hệ thống gợi ý món ăn sử dụng mô hình KNN, dữ liệu cần được chuẩn bị sao cho phản ánh chính xác mối quan hệ giữa các thành phần nguyên liệu, lượng dinh dưỡng (nutrition) và món ăn tương ứng. Mô hình KNN hoạt động dựa trên nguyên lý tìm kiếm các "hàng xóm gần nhất" trong không gian đặc trưng, do đó dữ liệu mong muốn phải có cấu trúc số hóa, định lượng và đồng nhất nhằm hỗ trợ việc tính toán khoảng cách giữa các điểm dữ liệu (ví dụ: Euclidean distance). Trong trường hợp này, người dùng sẽ nhập các thành phần hoặc lượng thành phần mong muốn (ví dụ: protein, chất béo, carbohydrate), và hệ thống sẽ đề xuất các món ăn phù hợp nhất dựa trên dữ liệu đã huấn luyện.

2. Đặc Điểm Dữ Liệu Mong Muốn

Dữ liệu lý tưởng để huấn luyện mô hình KNN trong hệ thống gợi ý món ăn cần đáp ứng các yêu cầu sau:

2.1. Cấu Trúc Dữ Liệu

- **Hình thức bảng (Tabular Data):** Dữ liệu nên được tổ chức dưới dạng bảng với các cột rõ ràng, mỗi cột đại diện cho một đặc trưng (feature) của món ăn.

2.2. Định Lượng và Chuẩn Hóa

- **Định lượng:** Các thành phần dinh dưỡng cần được biểu diễn dưới dạng số (ví dụ: Protein: 15g, Fat: 8g) thay vì định tính (cao, thấp). Điều này cho phép KNN tính toán khoảng cách chính xác giữa các món ăn và yêu cầu của người dùng.
- **Chuẩn hóa (Normalization):** Vì KNN nhạy cảm với độ lớn của các giá trị, dữ liệu cần được chuẩn hóa (ví dụ: Min-Max Scaling hoặc Z-score) để đảm bảo các thành phần như Protein và Calories có cùng thang đo, tránh trường hợp một đặc trưng có giá trị lớn lấn át các đặc trưng khác.

2.3. Tính Đầy Đủ và Sạch

- **Không có giá trị thiếu:** Các cột liên quan đến thành phần dinh dưỡng phải đầy đủ dữ liệu. Nếu thiếu, cần điền giá trị bằng trung bình, trung vị hoặc loại bỏ bản ghi đó.

- **Đồng nhất đơn vị:** Tất cả giá trị dinh dưỡng phải sử dụng cùng đơn vị (gram cho Protein/Fat/Carbs, kcal cho Calories) để đảm bảo tính nhất quán.

II. Nguồn dữ liệu

1. Nguồn Gốc Dữ Liệu

Dữ liệu là một phần của tập dữ liệu FoodRecSys-V1, được lấy từ nền tảng Kaggle. FoodRecSys-V1 là một tập dữ liệu công khai mô tả về các món ăn, bao gồm thông tin chi tiết về các công thức nấu ăn, nguyên liệu, hướng dẫn nấu, và giá trị dinh dưỡng của các món ăn.

Nguồn dữ liệu: foodRecSys-V1

2. Mô Tả Dữ Liệu

Dữ liệu được trình bày dưới dạng bảng với các cột chính như sau:

- recipe_id: Mã định danh duy nhất cho mỗi công thức (ví dụ: 222388, 240488, 218939).
- name: Tên món ăn (ví dụ: "Homemade Bacon", "Pork Loin, Apples, and Sauerkraut", "Foolproof Rosemary Chicken Wings").
- aver_rate: Điểm đánh giá trung bình của món ăn (ví dụ: 5.0, 4.764706, 4.571429).
- image_url: Đường dẫn đến hình ảnh của món ăn, thường từ trang Allrecipes.
- review_nums: Số lượng đánh giá cho món ăn (ví dụ: 3, 29, 12).
- ingredients: Danh sách nguyên liệu, được liệt kê dưới dạng chuỗi văn bản (ví dụ: "belly pork, paprika kosher salt ground b ..." cho món Homemade Bacon).
- cooking_directions: Hướng dẫn nấu ăn, được lưu dưới dạng chuỗi JSON (ví dụ: {'directions': 'uPrep\\n5 m\\nCook\\n2 h 45 m\\nRe...'}).
- nutrition: Thông tin dinh dưỡng, cũng được lưu dưới dạng chuỗi JSON (ví dụ: {'niacin': 0, 'hasCompleteData': False, 'u_name...'}).
- reviews: Thông tin đánh giá, bao gồm số lượng người theo dõi và điểm đánh giá (ví dụ: {'rating': 5, 'followersCount': 11, ...}).

3. Tại Sao Dữ Liệu Phù Hợp để Huấn Luyện Mô Hình KNN?

3.1. Các thành phần trong cột Nutition

- Thành phần dinh dưỡng có thể thấy được đầy đủ trong phần cột Nutition, phù hợp với yêu cầu đề ra để xây dựng hệ thống

3.2. Tính Đa Dạng của Nguyên Liệu và Món Ăn

- Cột ingredients chứa danh sách nguyên liệu đa dạng phản ánh nhiều loại món ăn từ các nền ẩm thực khác nhau. Điều này giúp hệ thống gợi ý có thể đáp ứng nhiều sở thích và yêu cầu của người dùng.

3.3. Số Lượng Bản Ghi Đủ Lớn

- Dữ liệu FoodRecSys-V1 trên Kaggle thường chứa hàng nghìn bản ghi. Số lượng lớn này đảm bảo rằng KNN có đủ "hàng xóm" để tìm kiếm và đưa ra gợi ý chính xác, tránh tình trạng thiếu dữ liệu dẫn đến kết quả không đáng tin cậy.

3.4. Hạn Chế và Cách Khắc Phục

Mặc dù dữ liệu phù hợp, vẫn có một số hạn chế cần lưu ý:

- Dữ liệu dinh dưỡng chưa đầy đủ: Một số bản ghi có hasCompleteData: False trong cột nutrition, nghĩa là thông tin dinh dưỡng không đầy đủ. Bạn cần xử lý bằng cách loại bỏ các bản ghi này hoặc điền giá trị thiếu (ví dụ: dùng trung bình của các món ăn tương tự).
- Chuẩn hóa: Các giá trị dinh dưỡng cần được chuẩn hóa (ví dụ: Min-Max Scaling) để đảm bảo các thành phần như Protein và Calories có cùng thang đo, tránh làm sai lệch kết quả của KNN.

II. Tiền xử lí dữ liệu

Phân Tích và Tiền Xử Lý Cột nutrition

1. Phân Tích Cột nutrition

Cột nutrition chứa thông tin dinh dưỡng của một món ăn, được lưu dưới dạng một chuỗi JSON với cấu trúc phức tạp.

Cấu Trúc Dữ Liệu

- hasCompleteData: Boolean (True/False), cho biết dữ liệu của thành phần này có đầy đủ hay không.
- name: Tên của thành phần (ví dụ: "Niacin Equivalents", "Sugars").
- amount: Giá trị định lượng của thành phần (ví dụ: 15.6016 cho niacin).
- percentDailyValue: Phần trăm giá trị hàng ngày (dựa trên chế độ ăn 2000 kcal/ngày).
- displayValue: Giá trị hiển thị (có thể làm tròn hoặc định dạng khác so với amount).
- unit: Đơn vị của thành phần (ví dụ: mg, g, kcal, IU, mcg).

- Tập trung vào percentDailyValue, vì nó biểu thị tỷ lệ phần trăm của giá trị dinh dưỡng hàng ngày (Daily Value - DV) mà món ăn cung cấp, dựa trên chế độ ăn 2000 kcal/ngày. Đây là một chỉ số quan trọng để phân tích lượng dinh dưỡng người dùng muốn nạp và phù hợp để huấn luyện mô hình.
-

2. Hướng Tiền Xử Lý Cột nutrition (Tập Trung vào percentDailyValue)

Bước 1: Tách JSON và Lấy percentDailyValue

- Mục tiêu: Chuyển đổi dữ liệu JSON thành các cột riêng biệt, chỉ lấy giá trị percentDailyValue của các thành phần dinh dưỡng cần thiết.
- Thành phần cần giữ: Tập trung vào các thành phần chính mà người dùng thường quan tâm:
    - calories
    - protein
    - fat
    - carbohydrates
    - (Tùy chọn) sugars, fiber, sodium, saturatedFat (nếu bạn muốn mở rộng).

Bước 2: Chuyển Đổi Định Dạng Chuỗi Thành Số

- Vấn đề: percentDailyValue được lưu dưới dạng chuỗi, cần chuyển thành số (float hoặc int).
- Hướng xử lý:
    - Chuyển các giá trị chuỗi thành số (float).
    - Xử lý các giá trị không hợp lệ (như '-' trong caloriesFromFat).

Bước 3: Xử Lý Dữ Liệu Thiếu hoặc Không Hợp Lệ

- Vấn đề: Một số bản ghi có thể thiếu percentDailyValue hoặc có giá trị không hợp lệ (NaN sau khi chuyển đổi).
- Hướng xử lý:
    - Điền giá trị thiếu: Nếu một bản ghi thiếu percentDailyValue cho một thành phần chính (calories, protein, fat, carbohydrates), có thể điền bằng giá trị trung bình của cột đó trong toàn bộ tập dữ liệu.
    - Loại bỏ bản ghi: Nếu một bản ghi thiếu quá nhiều thành phần chính, nên loại bỏ để đảm bảo chất lượng dữ liệu.

- Kiểm tra hasCompleteData: Dù hasCompleteData: False, percentDailyValue vẫn có giá trị (ví dụ: sodium có percentDailyValue: '104'). Do đó, ta có thể sử dụng các giá trị này mà không cần lo lắng về hasCompleteData.

VIII. Kết luận tạm thời

Tuần 2 sẽ tập trung vào tìm nguồn dữ liệu và tiền xử lý dữ liệu.