Trường Đại học Khoa học tự nhiên – Khoa Công nghệ thông tin.

Đồ án thực hành cuối kỳ

Nhập môn khoa học dữ liệu – Introduction to Data Science.

Nhóm 9
Tháng 11, 2024.

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN THỰC HÀNH CUỐI KÌ

Bộ môn: Nhập môn khoa học dữ liệu.
Tên đề tài:

"Phân tích mối quan hệ giữa các loại gia vị, nguyên liệu và khả năng kết hợp của chúng trong các món ăn".

STT nhóm: 9.

Thành viên:

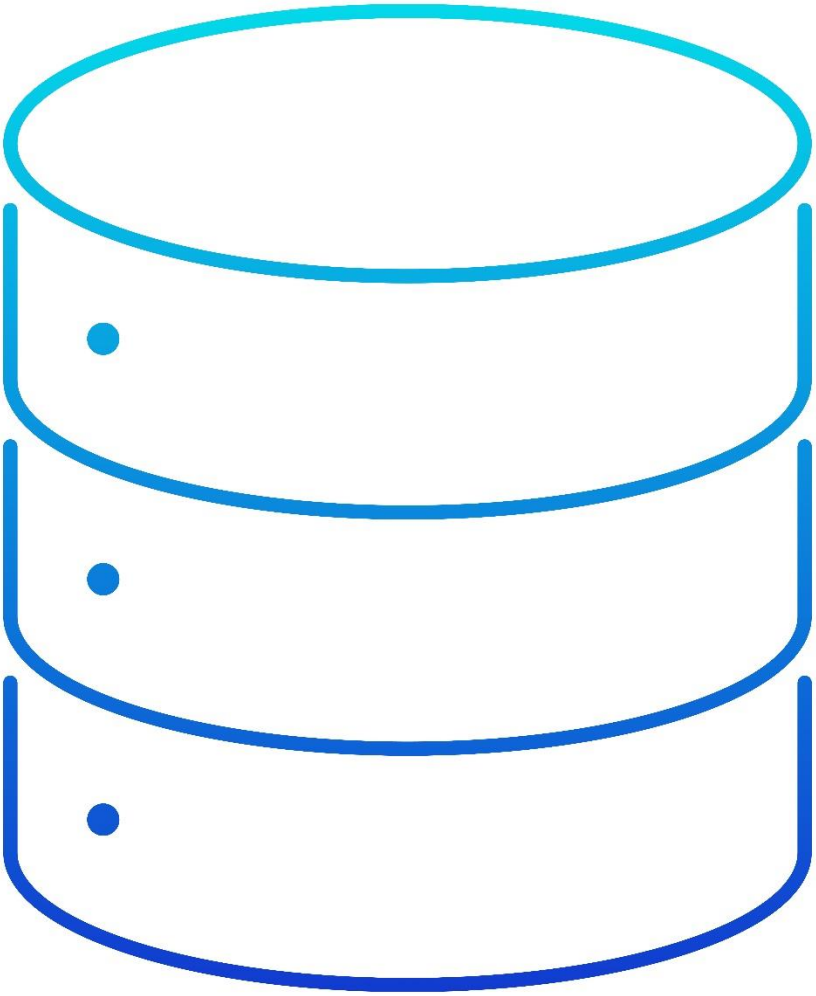
1. 22120384 – Nguyễn Đình Trí.
2. 22120398 – Vũ Hoàng Nhật Trường.
3. 22120412 – Nguyễn Anh Tường.
4. 22120424 – Nguyễn Ngọc Bảo Uyên.
5. 22120449 – Lê Nguyễn Huyền Vy.

Thông tin chung:

1. **Bộ môn:** Nhập môn khoa học dữ liệu.
2. **Giảng viên lý thuyết:** Thầy Lê Ngọc Thành.
3. **Giảng viên thực hành:** Thầy Lê Nhựt Nam.
4. **Mã lớp:** 22_21.
5. **STT nhóm:** 9.
6. **Danh sách thành viên:**
    a. 22120384 – Nguyễn Đình Trí.
    b. 22120398 – Vũ Hoàng Nhật Trường.
    c. 22120412 – Nguyễn Anh Tường.
    d. 22120424 – Nguyễn Ngọc Bảo Uyên.
    e. 22120449 – Lê Nguyễn Huyền Vy.
7. **Link github repository:** "Click here to go to our github repository."

MỤC LỤC

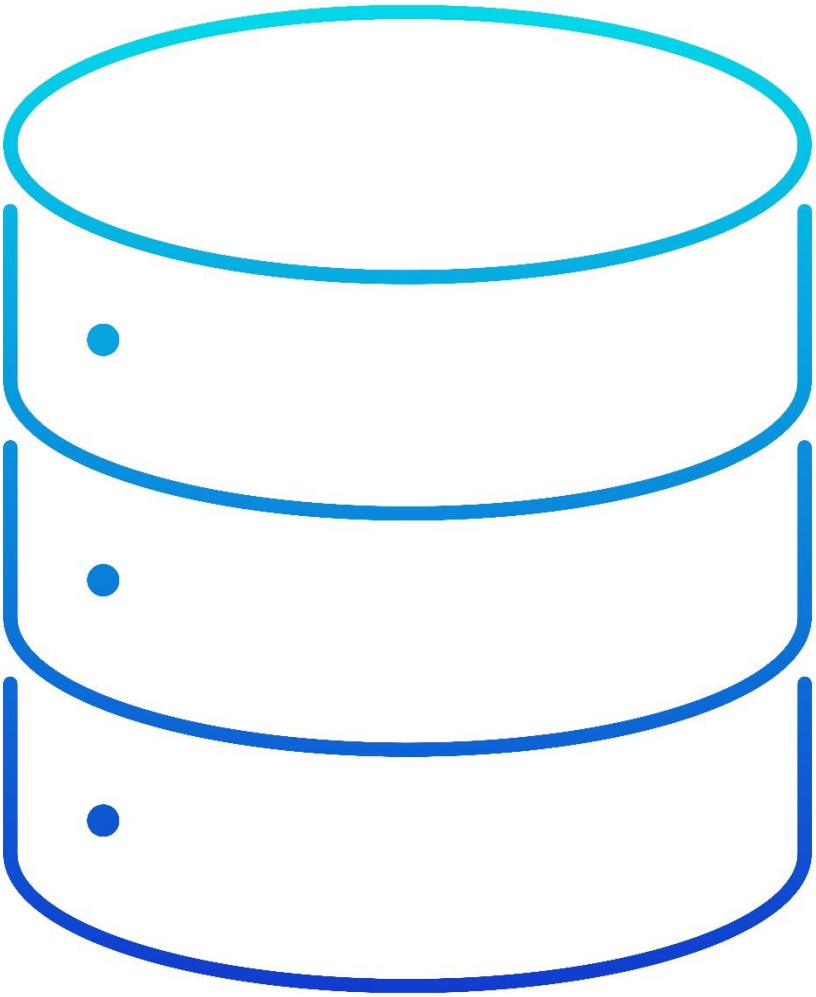
Section 0: *Bảng phân công công việc.*



I. Bảng phân công công việc:

Công việc	Người thực hiện	Tiến độ	Đóng góp	Note
Data Collection	Nguyễn Anh Tường	100%	90%	Các bạn chạy cào giúp vì bị trang web chặn.
	Nguyễn Đình Trí		2.5%	
	Hoàng Vũ Nhật Trường		2.5%	
	Nguyễn Ngọc Bảo Uyên		2.5%	
	Lê Nguyễn Huyền Vy		2.5%	
Data Preprocessing	Lê Nguyễn Huyền Vy	100%	95%	
	Nguyễn Anh Tường		5%	Thảo luận cách làm
Data Exploration	Nguyễn Đình Trí	0%		--
	Nguyễn Ngọc Bảo Uyên			
Data Modeling & Evaluation	Hoàng Vũ Nhật Trường	0%		--
	Nguyễn Anh Tường			

Section 1: *Data Collection.*



I. Các trường dữ liệu cần cào:

1. Tên bài viết:

Ví dụ:

- Title: Đổi Vị Với Phở Gà Trộn Lạ Miệng, Thơm Ngon Ăn Mãi Mà Không Thấy Ngán
- Title: Mì Ý Thịt Viên Sốt Cà Chua Nữ Hoàng Của Nền Ẩm Thực Nước Ý.
- Title: Khoai Tây Tẩm Gia Vị Nướng Kiểu Âu.

2. Tên của các nguyên liệu của món ăn đó:

Ví dụ:

Ingredients: ½ con gà ta 1 miếng gừng nhỏ (nạo vỏ, đập dập) 250ml nước 1 tsb muối 2 tbs xì dầu 2 tbs đường 1 tsp giấm tỏi 1 củ hành tây 500g bánh phở tươi 1 nắm giá Rau mùi, hành lá Hành phi Lạc rang

Ingredients: 6-8 củ khoai tây 1 tbs dầu olive ½ tsp muối ½ tsp tiêu xay ½ bột ớt paprika ½ tsp bột tỏi Hương thảo tươi

II. Định dạng của dữ liệu sau khi cào về:

Các món ăn hay bài viết khác nhau sẽ được phân tách với nhau bởi một dòng:

‘----------’

Ảnh minh họa:

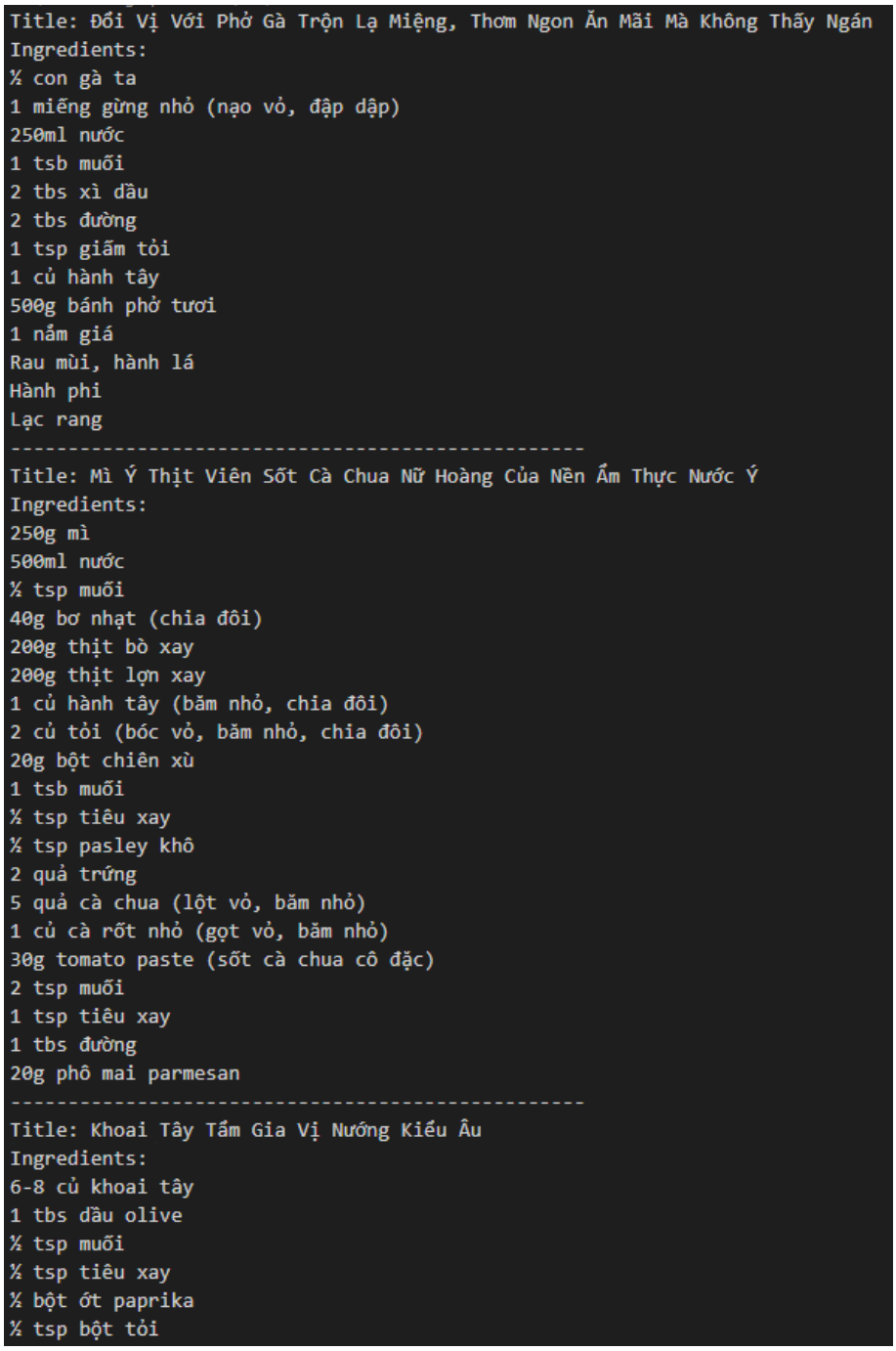


Figure 1. Ảnh minh họa cho tập dữ liệu sau khi cào.

III. Quá trình thực hiện:

1. Kokotaru.com:

a. Nhận xét: Sử dụng tương tác cuộn chuột để load trang và load bài viết.

b. Chi tiết các bước thực hiện cào dữ liệu từ Kokotaru:

Bước 1: Sử dụng selenium để giả lập trình duyệt chrome.

Bước 2: Sử dụng trình duyệt chrome với thao tác cuộn chuột qua tất cả các trang con của kokotaru.

Bước 3: Lấy hết tất cả các liên kết bài viết từ trang web.

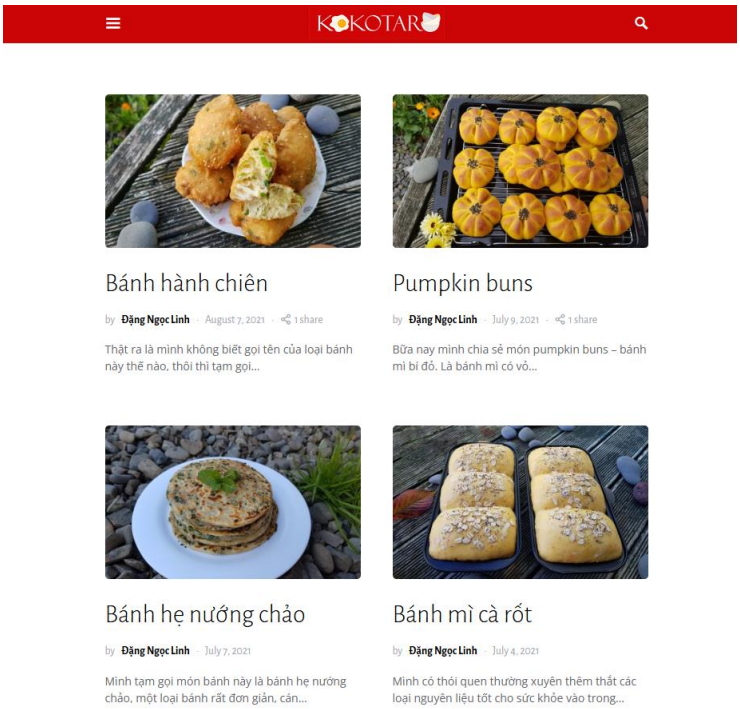


Figure 2. Ảnh trang web kokotaru.

Bước 4: Truy cập vào từng bài viết và lấy ra thành phần của các món ăn.

c. Một số vấn đề xảy ra:

- Người viết bài và code web viết ở quá nhiều dạng trình bày khác nhau.
- Các bài viết không có nguyên liệu.
- Các bài viết nguyên liệu được chèn sẵn vào trong tiêu đề.
- Các bài viết nguyên liệu được trộn vào trong dòng văn giới thiệu.
- Các bài viết sử dụng nguyên liệu A,B,C nhưng không dùng đến.

2. Kitchenart.com

a. Nhận xét: Các trang được phân chia theo dạng bình thường, gồm 38 trang. Trong mỗi trang sẽ có đường dẫn đến trang tiếp theo.

b. Chi tiết các bước cào dữ liệu:

Bước 1: Thực hiện lấy hết đường dẫn dẫn đến các trang chứa bài viết.

Bước 2: Thực hiện vào từng trang tổng hợp lấy ra hết các đường dẫn dẫn đến các bài viết có trong trang đó.

Bước 3: Thực hiện vào từng đường dẫn của bài viết để lấy dữ liệu.

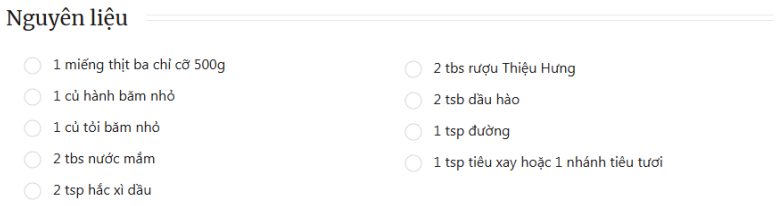


Figure 3. Ảnh mục nguyên liệu của kitchenart.com

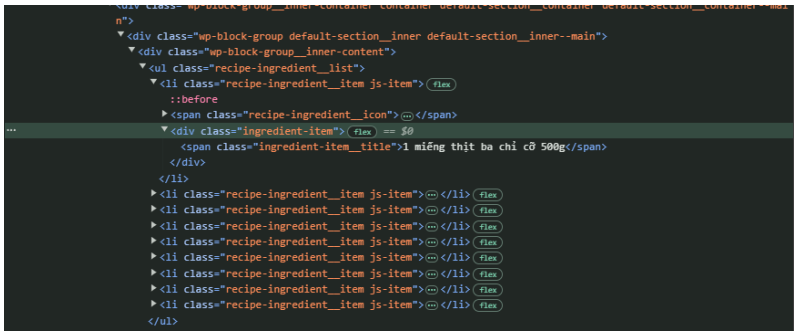


Figure 4. Mã HTML của phần nguyên liệu.

c. Một số vấn đề xảy ra:

- Trang web cấm truy cập với tốc độ rất nhanh và nhiều lần.
- Trung bình truy cập lần lượt 5-10 bài với tốc độ nhanh thì trang web đã phát hiện và tiến hành chặn truy cập. Đã thử truy cập 1 – 2 – 3 bài nhưng vẫn bị chặn.
- Thời gian chờ mở chặn rất lâu. Vẫn chưa có con số chính xác nhưng dao động trong khoảng từ 5 – 8 phút.

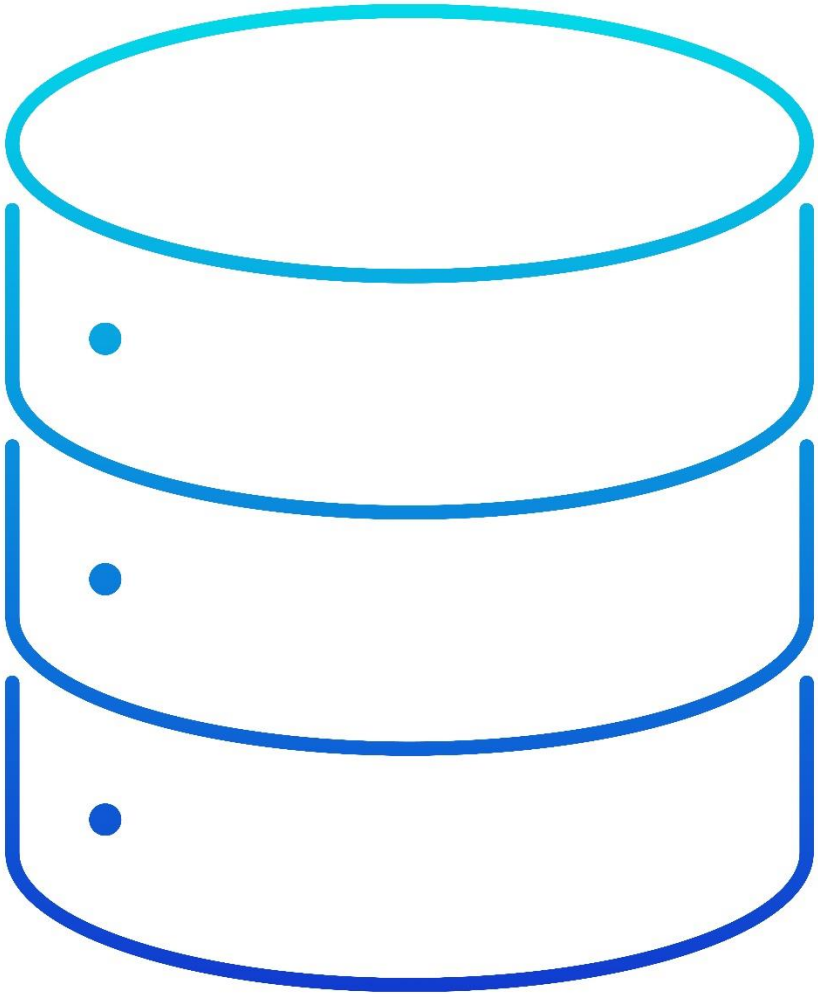
IV. Kết quả:

Kokotaru: 399 / 468 bài viết.

Kitchenart: 682 / 760 bài viết.

Tổng cộng: 1081 bài viết.

Section 2: *Data Cleaning and Normalizing*



I. Data Cleaning

1. Dịch tập tin

Để thuận tiện cho việc tiền xử lý và khám phá dữ liệu mà không phải quan tâm quá nhiều đến vấn đề xử lý unicode, các tập tin lưu dữ liệu được cào về từ hai website sẽ được dịch sang Tiếng Anh bằng Google Translate.

Trong quá trình dịch thuật, dữ liệu sẽ được kiểm tra lại để bảo đảm tính đồng nhất về bản dịch tên của các nguyên liệu.

2. Làm sạch dữ liệu:

Từng dòng dữ liệu sẽ được xử lý để đảm bảo chỉ giữ lại đúng tên nguyên liệu, lược bỏ các thông tin khác như lượng từ, tính từ, mô tả, cách sơ chế nguyên liệu đó.

a. Ví dụ:

Dữ liệu trước khi làm sạch

Ingredients: 6-8 củ khoai tây 1 tbs dầu olive ½ tsp muối ½ tsp tiêu xay ½ bột ớt paprika ½ tsp bột tỏi Hương thảo tươi

Dữ liệu sau khi làm sạch:

Ingredients: khoai tây dầu olive muối tiêu xay bột ớt paprika bột tỏi hương thảo tươi

b. Cách thực hiện

- Bởi vì từng các dòng nguyên liệu sẽ có các format khác nhau nên chúng ta không thể nào lọc lấy tên nguyên liệu bằng cách chỉ sử dụng những lệnh đơn giản như “bỏ đi kí số chỉ số lượng”, “lấy chuỗi con từ vị trí i đến vị trí j”…
- Cho nên, việc làm sạch nguyên liệu cần được hỗ trợ bởi những từ khóa, những ký tự đặc biệt.

Ví dụ:

- Những từ ngữ là lượng từ (half, clove, handful,…) hay đơn vị đo (g, kg, cup, bowl,…)
- Những từ ngữ chỉ tính chất hoặc cách sơ chế nguyên liệu đó (roasted, chopped, peeled,…)
- Những từ ngữ, kí hiệu là giới từ đánh dấu phần chú thích (*for* decorations, (),…)

• Các bước làm sạch một dòng dữ liệu:

- Chuẩn hóa Unicode, đưa chuỗi về kí tự thường.
- Loại bỏ nội dung trong ngoặc đơn.
- Tách thành các nguyên liệu khác nhau khi gặp các kí tự hoặc từ ngữ đóng vai trò phân cách.
- Loại bỏ kí tự không phải ASCII.
- Chuẩn hóa các phân số bị lỗi.
- Loại bỏ số lượng và đơn vị đo lường.
- Thay thế các kí tự không cần thiết còn sót lại thành khoảng trắng.
- Loại bỏ các thành phần là chú thích về cách sơ chế, tính chất của nguyên liệu.

-Chuẩn hóa về dạng số ít của nguyên liệu.

3. Định dạng của dữ liệu sau khi làm sạch:

Sau khi làm sạch, mỗi dòng nguyên liệu sẽ trả về một hoặc nhiều nguyên liệu đã được xử lý. Từ đó chuyển đến bước tiếp theo để tiến hành chuẩn hóa dữ liệu.

II. Data Normalizing

1. “Gộp” dữ liệu

Sau khi làm sạch, mỗi món ăn sẽ có tên món ăn và một danh sách chứa các nguyên liệu của món ăn đó.

Nhiệm vụ của chúng ta là tạo ra một danh sách chứa tất cả các nguyên liệu phân biệt có trong tất cả những món ăn được cào về từ website ban đầu. Ta sẽ tiến hành tạo ra một dataframe với từng dòng biểu thị cho từng món ăn, từng cột biểu thị cho từng nguyên liệu.

2. Chuẩn hóa dữ liệu

Sau khi dữ liệu được đọc từ tập tin, ta đưa thông tin của từng món ăn bao gồm tên món ăn và danh sách nguyên liệu vào một danh sách (list). Mỗi phần tử trong đó sẽ có kiểu dictionary với key ‘title’ lưu tên món ăn và key ‘ingredients’ lưu danh sách các nguyên liệu.

Các bước thực hiện:

- Đọc file và xử lý từng món ăn (dựa vào dòng phân cách 50 kí tự ‘-‘)
- Lưu tên món ăn và danh sách nguyên liệu đã được làm sạch vào combined_data

```
combined_data.append({
```

```
                'title': title,
                'ingredients': clean_ingredients
            })
```

- Tạo dataframe từ combined_data
- Lấy danh sách các nguyên liệu phân biệt trong dataframe và sắp xếp lại theo thứ tự bảng chữ cái.
- Tạo một dataframe nhị phân gồm cột "Name of dish" và các cột còn lại là tên các nguyên liệu có trong danh sách nguyên liệu phân biệt.
- Điền giá trị vào dataframe: Nếu món ăn nằm ở dòng i có sử dụng nguyên liệu ở cột j thì vị trí tương ứng đó có giá trị bằng 1, ngược lại bằng 0.

III. Kết quả

Sau khi thực hiện làm sạch và chuẩn hóa dữ liệu, ta kiểm tra xem số lượng dòng của dataframe ta tạo ra có khớp với số lượng món ăn đã cào được ở phần Data Collection hay không.

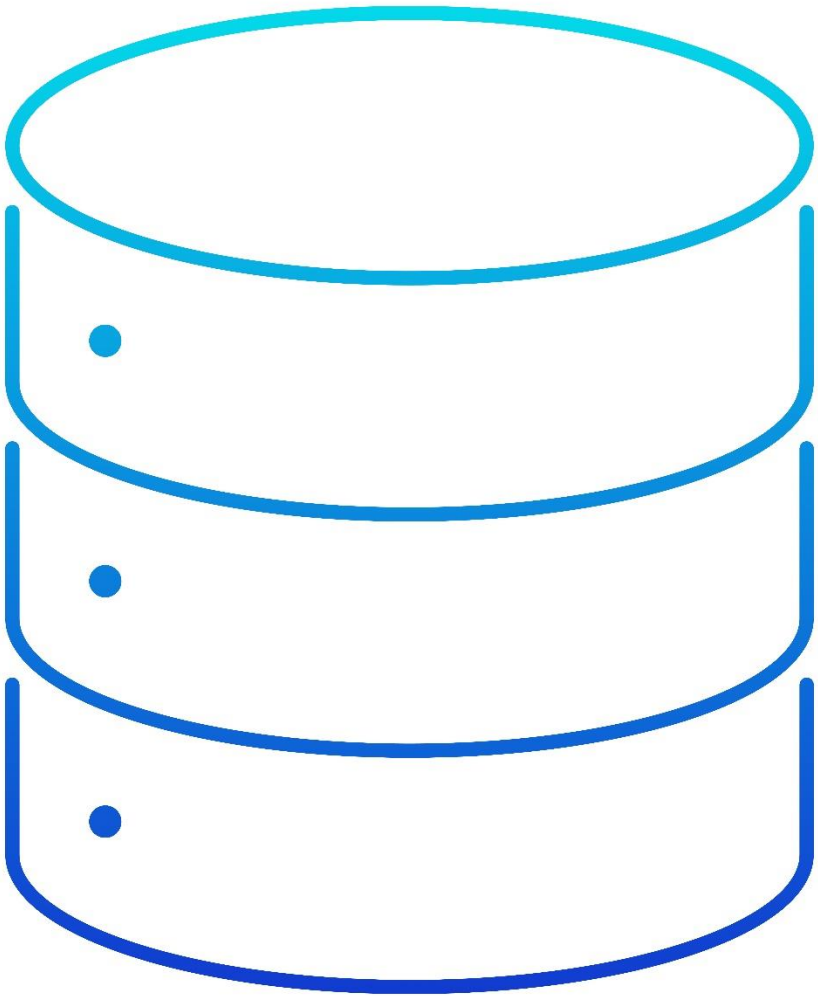
```
binary_df, all_ingredients = normalize_recipes_to_dataframe (file_paths,units,key_words,black_list,lemmatizer)

print(f"Number of dishes: {binary_df.shape[0]}")
print(f"Number of ingredients: {binary_df.shape[1]}")
```

Kết quả thu được: Number of dishes: 682 (khớp với số lượng món ăn trong file dữ liệu được cào về từ web Kitchenart).

Đối với website Kokotaru, bởi vì cấu trúc đặc biệt của web đã dẫn đến việc trong một số trường hợp, kết quả trả về từ việc cào dữ liệu không phải là nguyên liệu của món ăn. Trong một số trường hợp khác, tên nguyên liệu không được trình bày thành các dòng phân biệt mà được biểu thị trong một đoạn văn miêu tả, cho nên phải mất thời gian lâu hơn cũng như phương pháp phức tạp hơn để giải quyết vấn đề này.

Phần 3: *Kế hoạch trong tương lai.*



I. Các bước thực hiên:

Thực hiện tiếp từ bước 4: Khám phá, đánh giá, trực quan và phân tích dữ liệu.

Bước 5: Xây dựng tập huấn luyện, xây dựng mô hình, và đánh giá mô hình.

Bước 6: Trực quan hóa kết quả và báo cáo kết quả.

Bước 7: Nhận xét đánh giá và duy trì các kết quả theo kỳ vọng.

II. Kết quả kỳ vọng:

1. Phân tích dữ liệu:

a. Phân tích tần suất sử dụng gia vị, nguyên liệu:

- Xác định tần suất xuất hiện của mỗi loại gia vị, nguyên liệu trong các món ăn.
- Từ đó, có thể tìm ra các loại gia vị, nguyên liệu phổ biến nhất được sử dụng trong nhiều món ăn, cũng như các gia vị, nguyên liệu ít xuất hiện hơn.

b. Phân nhóm món ăn theo loại gia vị, nguyên liệu chính:

- Dựa trên các gia vị, nguyên liệu, có thể nhóm các món ăn lại theo loại gia vị, nguyên liệu chính (ví dụ: nhóm các món có ớt, gừng, hay tỏi làm gia vị chính).
- Điều này giúp ta hiểu được sự liên quan giữa các món ăn dựa trên sự tương đồng về gia vị, nguyên liệu.

c. Phân tích tương quan giữa các gia vị, nguyên liệu:

- Phân tích xem các gia vị, nguyên liệu thường xuyên đi kèm với nhau. Ví dụ, nếu tỏi và ớt thường xuất hiện cùng nhau, suy ra giữa có mối quan hệ trong việc tạo hương vị.
- Phân tích xem các loại gia vị, nguyên liệu nào thường xuyên không đi kèm với nhau. Nguyên nhân?

d. Đánh giá độ phức tạp của món ăn dựa trên số lượng gia vị, nguyên liệu:

- Phân loại và đánh giá các món ăn dựa trên độ phức tạp của gia vị, nguyên liệu.
- Có thể có 5 mức độ đánh giá tương ứng với số ngôi sao.

2. Phát triển mô hình (KỲ VỌNG):

a. Mô hình gợi ý gia vị thay thế, nguyên liệu:

Phát triển mô hình gợi ý gia vị, nguyên liệu thay thế dựa trên các món ăn tương tự hoặc các món sử dụng các gia vị, nguyên liệu gần giống nhau.

Mô hình này có thể sử dụng Word2Vec hoặc cosine similarity để tìm ra các gia vị, nguyên liệu thay thế phù hợp trong cùng một nhóm.

b. Mô hình phân cụm món ăn theo gia vị, nguyên liệu:

Sử dụng các thuật toán như K-Means Clustering để phân cụm món ăn dựa trên các loại gia vị, nguyên liệu. Mô hình sẽ giúp chia món ăn thành các nhóm khác nhau dựa trên sự tương đồng về thành phần gia vị, nguyên liệu.

c. Mô hình đề xuất món ăn dựa trên gia vị, nguyên liệu có sẵn:

Phát triển hệ thống đề xuất món ăn dựa trên các gia vị, nguyên liệu người dùng có sẵn.

Sử dụng các kỹ thuật như collaborative filtering hoặc content-based filtering để đề xuất những món ăn có thể nấu dựa trên các gia vị, nguyên liệu mà người dùng nhập vào.