**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO TIỂU LUẬN**

**MÔN: KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI:**

**TẬP PHỔ BIẾN VÀ ỨNG DỤNG MINH HỌA**

**TP. HỒ CHÍ MINH, 27 THÁNG 02 NĂM 2025**

**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỀ TÀI:**

**TẬP PHỔ BIẾN VÀ ỨNG DỤNG MINH HỌA**

**Giảng viên hướng dẫn:** ThS. Trần Như Ý

**Sinh viên thực hiện:**

1. Cao Thiên Chi 2001220536
2. Trương Mỹ Hoa 2001221445
3. Lương Tấn Tài 2001224242

**TP. HỒ CHÍ MINH, 27 THÁNG 02 NĂM 2025**

LỜI CẢM ƠN

Để hoàn thành được bài tiểu luận này, nhóm chúng em xin chân thành cảm ơn Ban Giám hiệu, các khoa, phòng và quý thầy, cô của trường đại học Công Thương TP. Hồ Chí Minh người đã tận tình giúp đỡ và tạo điều kiện cho em trong quá trình học tập. Đặc biệt, nhóm xin gửi lời cảm ơn sâu sắc đến cô Trần Như Ý - người đã trực tiếp giảng dạy và hướng dẫn chúng em thực hiện bài tiểu luận này bằng tất cả lòng nhiệt tình và sự quan tâm sâu sắc.

Trong quá trình thực hiện bài tiểu luận này, do hiểu biết còn nhiều hạn chế nên bài làm khó tránh khỏi những thiếu sót. Nhóm rất mong nhận được những lời góp ý của quý thầy cô để bài tiểu luận ngày càng hoàn thiện hơn.

**Nhóm chúng em xin chân thành cảm ơn!**

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc191565449)

[MỤC LỤC ii](#_Toc191565450)

[DANH MỤC HÌNH ẢNH iv](#_Toc191565451)

[DANH MỤC TỪ VIẾT TẮT v](#_Toc191565452)

[BẢNG PHÂN CÔNG NHIỆM VỤ vi](#_Toc191565453)

[PHẦN NỘI DUNG 1](#_Toc191565454)

[Chương 1. GIỚI THIỆU ĐỀ TÀI 1](#_Toc191565455)

[I. Định nghĩa vấn đề 1](#_Toc191565456)

[II. Phạm vi của đề tài 1](#_Toc191565457)

[III. Mục tiêu của đề tài 1](#_Toc191565458)

[Chương 2. CÁC ĐỊNH NGHĨA VỀ TẬP PHỔ BIẾN 2](#_Toc191565459)

[I. Bài toán khai thác tập phổ biến 2](#_Toc191565460)

[II. Các khái niệm trong bài toán khai thác tập dữ liệu 2](#_Toc191565461)

[1. Các khái niệm cơ bản 2](#_Toc191565462)

[2. Độ hỗ trợ (Support) và tập phổ biến (Frequent Itemsets) 5](#_Toc191565463)

[3. Tập phổ biến tối đại (Maximal Frequent Itemsets) 7](#_Toc191565464)

[4. Tập phổ biến đóng (Closed Frequent Itemsets) 8](#_Toc191565465)

[III. Mối liên hệ giữa các tập phổ biến 9](#_Toc191565466)

[IV. Các thuật toán khai thác tập phổ biến 10](#_Toc191565467)

[Chương 3. ỨNG DỤNG MINH HỌA 11](#_Toc191565468)

[I. Mục tiêu 11](#_Toc191565469)

[II. Thuật toán RELim (Recursive Elimination Algorithm) 11](#_Toc191565470)

[1. Định nghĩa 11](#_Toc191565471)

[2. Ý tưởng 11](#_Toc191565472)

[3. Thuật toán 11](#_Toc191565473)

[4. Ví dụ 12](#_Toc191565474)

[5. Lưu ý 15](#_Toc191565475)

[III. Ứng dụng tìm tập phổ biến không theo thuật toán 15](#_Toc191565476)

[1. Giao diện 15](#_Toc191565477)

[2. Cách sử dụng 15](#_Toc191565478)

[IV. Ứng dụng tìm tập phổ biến theo thuật toán Relim 17](#_Toc191565479)

[1. Giao diện 17](#_Toc191565480)

[2. Code 18](#_Toc191565481)

[TÀI LIỆU THAM KHẢO 20](#_Toc191565482)

DANH MỤC HÌNH ẢNH

[Hình 2.1. Ví dụ về CSDL giao dịch 4](#_Toc191559959)

[Hình 2.2: CSDL sau khi biển đổi thành dạng nhị phân. 4](#_Toc191559960)

[Hình 2.3: Ví dụ về tập phổ biến 5](#_Toc191559961)

[Hình 2.4: CSDL sau khi biến đổi về dạng nhị phân 6](#_Toc191559962)

[Hình 2.5: Ví dụ về tập phổ biến tối đại 8](#_Toc191559963)

[Hình 2.6: Ví dụ về tập phổ biến đóng 8](#_Toc191559964)

[Hình 2.6: Quan hệ giữa các tập phổ biến 9](#_Toc191559965)

[Hình 4.1: Ví dụ thuật toán RELim 12](#_Toc191559966)

[Hình 4.2: CSDL sau khi biến đổi sang dạng nhị phân 12](#_Toc191559967)

[Hình 4.3: Dữ liệu sau giai đoạn 1 13](#_Toc191559968)

[Hình 4.4: Dữ liệu sau giai đoạn 2 14](#_Toc191559969)

[Hình 4.4: Dữ liệu sau giai đoạn 3 14](#_Toc191559970)

[Hình 4.5: Giao diện ứng dụng 15](#_Toc191559971)

DANH MỤC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| **Từ viết tắt** | **Dịch nghĩa** |
| CSDL | Cơ sở dữ liệu |
| TID | Mã giao dịch |
| Supp | Độ hỗ trợ |
| Minsupp | Ngưỡng hỗ trợ tối thiểu |

BẢNG PHÂN CÔNG NHIỆM VỤ

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên Sinh Viên** | **MSSV** | **Nhiệm Vụ** | **Tỉ Lệ  Hoàn Thành** |
| Cao Thiên Chi | 2001220536 | Word + PPT | 100% |
| Trương Mỹ Hoa | 2001221445 | Code + Word | 100% |
| Lương Tấn Tài | 2001224242 | Code + PPT | 100% |

PHẦN NỘI DUNG

# GIỚI THIỆU ĐỀ TÀI

## Định nghĩa vấn đề

Các cửa hàng đối mặt với những vấn đề sau:

* Khách hàng sẽ bị thu hút bởi những sản phẩm nào?
* Khách hàng thường mua sắm những vật dụng nào?
* Các kệ hàng cần được xếp như thế nào để khách hàng bị thu hút nhiều hơn?
* Cần sắp xếp hàng hóa để tối ưu thời gian mua sắm của khách hàng?

## Phạm vi của đề tài

Đề tài tập trung vào các định nghĩa liên quan đến tập phổ biến như:

* Hạng mục.
* Tập các hạng mục.
* Giao dịch.
* CSDL giao dịch.
* Độ phổ biến.
* Tập phổ biến.
* Tập phổ biến đóng.
* Tập phổ biến tối đại.

## Mục tiêu của đề tài

* Giới thiệu các kiến thức cơ bản về tập phổ biến.
* Xây dựng ứng dụng minh họa cho ứng dụng của tập phổ biến trong thực tế.

# CÁC ĐỊNH NGHĨA VỀ TẬP PHỔ BIẾN

## Bài toán khai thác tập phổ biến

Bài toán “Khai thác tập phổ biến” là bài toán rất quan trọng trong lĩnh vực khai phá dữ liệu vì nó vạch ra được tính chất ẩn, mô tả được những tính chất quan trọng trong dữ liệu mà chúng ta khó nhận ra trong phân tích thông thường. [3]

Bài toán khai thác tập dữ liệu còn là nền tảng của nhiều nhiệm vụ khai thác dữ liệu khác như:

* Phân tích luật kết hợp, mối tương quan.
* Mẫu tuần tự, cấu trúc.
* Phân tích dữ liệu không gian, dữ liệu đa phương tiện (văn bản, hình ảnh, âm thanh,..), dữ liệu dựa theo thời gian (ví dụ: dữ liệu thị trường chứng khoán thay đổi liên tục theo thời gian).
* Phân loại: phân loại dựa trên luật kết hợp.
* Phân tích nhóm: gom nhóm dự trên mẫu phổ biến.

Bài toán “Khai thác tập phổ biến” được đưa ra vào năm 1993 và được tiến hành phân tích trên một CSDL bán hàng hay còn gọi là CSDL giao dịch.

**CSDL giao dịch** là cơ sở dữ liệu được thu thập từ các dữ liệu bán hàng ở siêu thị hoặc các trung tâm thương mại. Trong đó, người ta sẽ thống kê xem trong một lần mua hàng, khách hàng sẽ mua những sản phẩm gì.

**Mục đích:** tìm các hiện tượng thường xuyên xảy ra trong tập dữ liệu.

* Những sản phẩm nào thường được bán cùng nhau trong cửa hàng?
* Khách hàng thường mua gì tiếp theo sau khi mua máy tính?

**Ứng dụng:**

* Áp dụng trong phân tích cơ sở dữ liệu bán hàng.
* Mở rộng sang quảng cáo, thiết kế chiến dịch bán hàng, chuỗi DNA,..

## Các khái niệm trong bài toán khai thác tập dữ liệu

### Các khái niệm cơ bản

* **Mẫu phổ biến** là mẫu (tập các hạng mục, cấu trúc con, đồ thị con,… ) xuất hiện thường xuyên trong cơ sở dữ liệu.

Khái niệm này được đề cập trong bối cảnh nghiên cứu bởi **Agrawal, Imielinski, Swami** vào năm 1993 và sau đó được công bố trong bài báo có tiêu đề **"Fast Algorithms for Mining Association Rules"** năm 1994.

* **Hạng mục (Item)** là một định nghĩa dùng để chỉ một mặt hàng trong giỏ hàng hay một thuộc tính trong cơ sở dữ liệu.
* **Tập các hạng mục (Itemset):** là tập chỉ tất cả các mặt hàng có trong cửa hàng hay tập tất cả các thuộc tính có trong CSDL.

Ký hiệu: I

* **Tập k hạng mục (k-itemset):** CSDL ban đầu sẽ có tập I mặt hàng, chỉ tất cả các mặt hàng có trong cửa hàng nhưng người mua không mua hết tất cả các mặt hàng mà chỉ một tập nhỏ các mặt hàng từ tập I, tập nhỏ các mặt hàng đó gọi là tập k hạng mục.

**Ví dụ:** Trong giỏ của khách hàng có 5 mặt hàng thì k =5 (5-itemset).

* **Giao dịch (Transation)**: là tập các hạng mục được mua trong một giỏ và giao dịch đó khi được lưu trong CSDL sẽ có một mã giao dịch TID.

Ký hiệu: t.

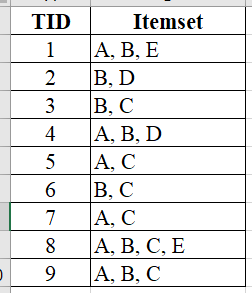
**Giao dịch t**: tập các hạng mục sao cho tI (t là tập con của I).

* **CSDL giao dịch** hay **tập chứa các giao dịch**: là dữ liệu chứa các giao dịch, gồm có n bản ghi dữ liệu. [4]

Ký hiệu: D. **CSDL D =** {t1, t2,…, tn}.

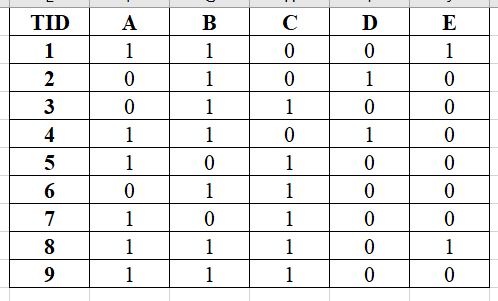
Trong quá trình phân tích CSDL, CSDL giao dịch sẽ được biến đổi về dạng nhị phân để dễ dàng quan sát, xem xét.

**Để minh họa cho các khái niệm, ta lấy ví dụ CSDL với các giao dịch sau:**



Hình 2.1. Ví dụ về CSDL giao dịch

CSDL trên có thể được chuyển về dạng nhị phân như sau:



Hình 2.2: CSDL sau khi biển đổi thành dạng nhị phân.

* **Hạng mục (item):** mặt hàng A, B, C, D, E.
* **Tập các hạng mục (itemset):** danh sách các hạng mục trong giỏ hàng .

I = {A, B, C, D , E}.

* **Mẫu phổ biến (frequent item):** là mẫu xuất hiện thường xuyên trong tập dữ liệu như {A, C} xuất hiện khá nhiều trong các giao dịch.
* **Tập k-hạng mục (k-itemset):**
  + k =2: danh sách cặp sản phẩm (2-itemset) như {{B, D}, {B, C}}.
  + k=3: danh sách 3 sản phẩm (3-itemset) đi kèm như {A, B, E}.
  + k=4: danh sách 4 sản phẩm (4-itemset) đi kèm như {A, B, C, E}.

### Độ hỗ trợ (Support) và tập phổ biến (Frequent Itemsets)

* **Giao dịch t chứa X** nếu X là tập các hạng mục trong I và X t

(X là tập con của t).

Chẳng hạn ta có giao dịch t={bánh mì, sữa chua, ngũ cốc} và tập X={Bánh mì, sữa chua}, ta có thể nói giao dịch t chứa X.

* **Độ độ hỗ trợ (Support)** của tập các hạng mục X trong CSDL D là tỷ lệ giữa số các giao dịch chứa X trên tổng số các giao dịch trong D.

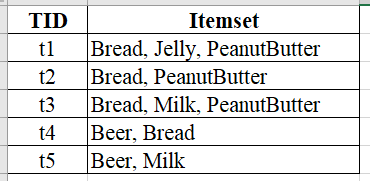
**Supp(X) = count(X) /**

* **Tập các hạng mục phổ biến S** hay **Tập phổ biến (Frequent Itemsets)**: là tập các hạng mục có độ hỗ trợ thỏa mãn độ hỗ trợ tối thiểu ***minsupp*** (do người dùng xác định, chẳng hạn minsupp=40% hoặc minsupp = 5 lần). [2]

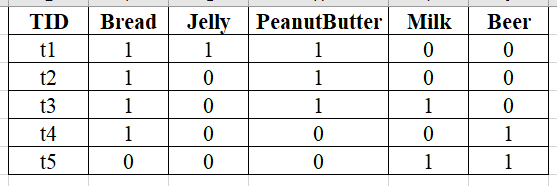
Nếu **supp(S) >= minsupp** thì **S là tập phổ biến**

* **Ví dụ tập phổ biến**

Với ***minsupp = 60%***

****

Hình 2.3: Ví dụ về tập phổ biến



Hình 2.4: CSDL sau khi biến đổi về dạng nhị phân

Ta có: I = {Bread, Jelly, PeanutButter, Milk, Beer}.

Xét X = {Bread, PeanutButter}, ta có:

* count(X) = 3 (Bread và PeanutButter cùng xuất hiện ở giao dịch t1, t2 và t3)
* |D| = 5 (tổng số giao dịch là 5)

Suy ra: Độ hỗ trợ của X: supp(X) = = 0.6 = 60% = minsupp

Từ đó ta có thể kết luận: X là một tập phổ biến.

**Trường hợp 1: X1 và X2 là tập con của X.**

Với X1 = { Bread }, ta có:

* count(X1) = 4 (Bread xuất hiện ở các giao dịch t1, t2, t3 và t4).
* |D| = 5 (tổng số giao dịch là 5).

Suy ra: Độ hỗ trợ của X1: supp(X1) = = 0.8 = 80% > minsupp (60%).

Từ đó ta có thể kết luận: X1 là một tập phổ biến.

Với X2 = { PeanutButter }, ta có:

* count(X2) = 4 (PeanutButter xuất hiện ở các giao dịch t1, t2, và t3).
* |D| = 5 (tổng số giao dịch là 5).

Suy ra: Độ hỗ trợ của X2: supp(X) = = 0.6 = 60% = minsupp (60%).

Từ đó ta có thể kết luận: X2 là một tập phổ biến.

**Trường hợp 2: X3 và X4 không là tập con của X.**

Với X3 = { Milk }, ta có:

* count(X3) = 2 (Milk xuất hiện ở giao dịch t3 và t5).
* |D| = 5 (tổng số giao dịch là 5).

Suy ra: Độ hỗ trợ của X3: supp(X) = = 0.4 = 40% < minsupp (60%).

Từ đó ta có thể kết luận: X3 không là một tập phổ biến.

Với X4 = { Milk, Bread }, ta có:

* count(X4) = 1 (Milk và Bread chỉ xuất hiện ở giao dịch t3).
* |D| = 5 (tổng số giao dịch là 5).

Suy ra: Độ hỗ trợ của X4: supp(X) = = 0.2 = 20% < minsupp (60%).

Từ đó ta có thể kết luận: X4 không là một tập phổ biến.

* **Nhận xét:** Với X là tập phổ biến:
* Trường hợp 1: X1 và X2 là tập con của tập X và là tập phổ biến.
* Trường hợp 2: X3 và X4 không là tập con của tập X và không là tập phổ biến.
* **Tính chất của tập phổ biến:**
* Tất cả các tập con của tập phổ biến đều là tập phổ biến.
* Nếu một tập không phải là tập phổ biến thì tất cả các tập bao của nó đều không phải là tập phổ biến.

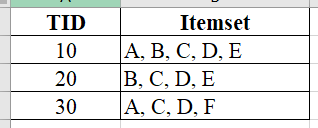
### Tập phổ biến tối đại (Maximal Frequent Itemsets)

Trong trường hợp minsupp quá cao thì số lượng tập phổ biến quá ít hoặc không có, còn minsupp thấp quá thì số lượng tập phổ biến quá cao dẫn đến khó phân tích, xem xét.

Năm 1998, người ta đưa ra khái niệm Tập phổ biến tối đại nhằm một bước nào đó hạn chế lại số lượng tập phổ biến.

**Tập phổ biến tối đại** là tập phổ biến và không tồn tại tập nào bao nó là phổ biến.

**Ví dụ:**



Hình 2.5: Ví dụ về tập phổ biến tối đại

Với minsupp=2 (số lần xuất hiện của Item).

Ta có các tập phổ biến là { {A}, {B}, {C}, {D}, {E},

{B, C}, {C, D},

{A, C, D}, {B, C, D}, {C, D, E},

{B, C, D, E} }.

Từ các tập phổ biến trên, ta có {A, C, D} và {B, C, D, E} là các tập phổ biến tối đại, vì chúng là tập lớn nhất, không còn tập phổ biến nào bao chúng.

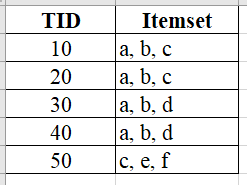
Giải thích thêm: Tập {A, B, C, D, E} là tập bao của tập {B, C, D, E} nhưng nó không phải là tập phổ biến nên {A, B, C, D, E} không là tập phổ biến tối đại.

### Tập phổ biến đóng (Closed Frequent Itemsets)

Người ta thấy rằng ở khái niệm tập phổ biến tối đại, số lượng tập phổ biến bị lược bỏ quá nhiều, nên năm 1999 người ta đưa ra khái niệm là Tập bao phổ biến hay Tập phổ biến đóng là một hình thức nén tập phổ biến lại dưới dạng có mất mát thông tin nhưng tốt hơn tập phổ biến tối đại, giúp giữ lại được nhiều số lượng tập phổ biến hơn.

**Tập phổ biến đóng** là tập phổ biến và không tồn tại tập nào bao nó có cùng độ hỗ trợ. Tập bao nó có thể là phổ biến nhưng phải có độ hỗ trợ nhỏ hơn nó.

**Ví dụ:**



Hình 2.6: Ví dụ về tập phổ biến đóng

Với minsupp = 2.

Ta có các tập phổ biến: { {a}, {b}, // supp = 4

{c}, // supp = 3

{d}, // supp = 2

{a, b}, // supp = 4

{a, c}, {a, d}, {b, c}, {b, d}, // supp = 2

{a, b, c}, {a, b, d} } // supp = 2

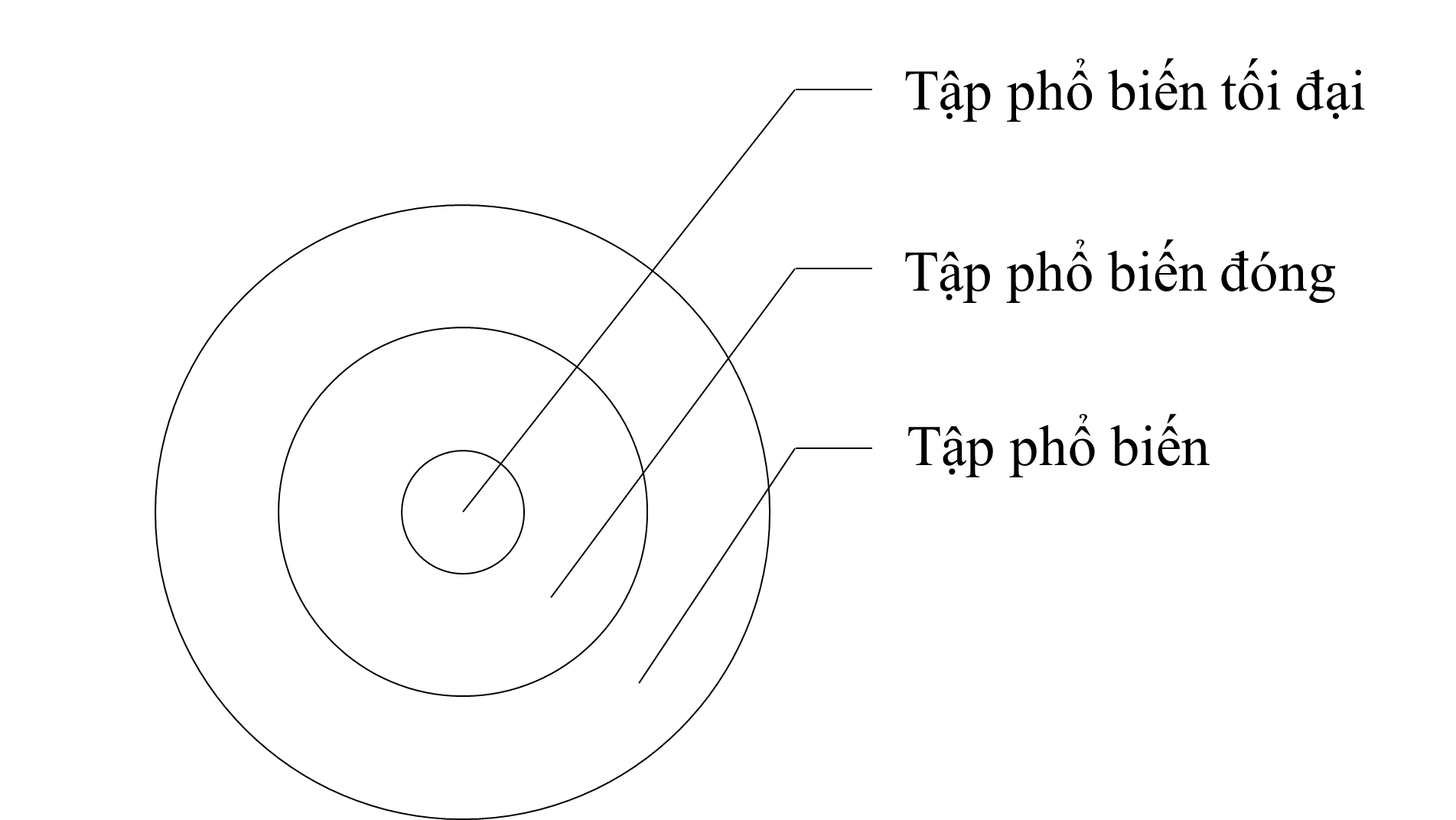
Từ các tập phổ biến trên, ta có {c}, {a, b}, {a, b, c}, {a, b, d} là các tập phổ biến đóng.

Giải thích:

* {c} là tập con của {a, c}, {b, c}, {a, b, c} và có độ hỗ trợ lớn hơn (3 > 2).
* {a, b} là tập con của {a, b, c}, {a, b, d} và có độ hỗ trợ lớn hơn (4 > 2).
* {a, b, c} và {a, b, d} không có tập nào bao nó.

## Mối liên hệ giữa các tập phổ biến

**Tập phổ biến ⊇ Tập phổ biến đóng ⊇ Tập phổ biến tối đại.**



Hình 2.6: Quan hệ giữa các tập phổ biến

## Các thuật toán khai thác tập phổ biến

Có nhiều thuật toán khai thác tập phổ biến như: [8]

* Thuật toán Apriori.
* Thuật toán FP\_Growth (Frequent Pattern Growth).
* Thuật toán Eclat (Equivalence Class Transformation).
* Thuật toán dựa trên Tidset.
* Thuật toán RELim (Recursive Elimination).
* Thuật toán Charm.
* Thuật toán LFIMiner\_ALL (Longest Frequent Itemset Mining).[5]
* Thuật toán MaxLFI (Maximum Length Frequent Itemset Mining). [5]

# ỨNG DỤNG MINH HỌA

## Mục tiêu

Dựa trên CSDL hóa đơn mua sắm và thuật toán Relim, ứng dụng sẽ đề xuất những sản phẩm liên quan đến sản phẩm mà người dùng tìm kiếm và gợi ý cho nhân viên những sản phẩm mà cửa hàng bán chạy nhất.

## Thuật toán RELim (Recursive Elimination Algorithm)

### Định nghĩa

**RELim (Recursive Elimination Algorithm)** là một thuật toán để khám phá các tập mục (nhóm mục) xuất hiện thường xuyên trong cơ sở dữ liệu giao dịch ( tập mục thường xuyên ). Một tập mục thường xuyên là một tập mục xuất hiện trong ít nhất minsup giao dịch từ cơ sở dữ liệu giao dịch, trong đó minsup là một tham số do người dùng cung cấp. [6]

Recursive Elimination là một thuật toán để tìm các tập mục thường xuyên, được lấy cảm hứng mạnh mẽ từ thuật toán FP-growth và rất giống với thuật toán H-mine.

Thuật toán này thực hiện công việc của mình mà không cần cây tiền tố hoặc bất kỳ cấu trúc dữ liệu phức tạp nào khác, xử lý trực tiếp các giao dịch. [7]

Điểm mạnh chính của thuật toán này không phải là tốc độ (mặc dù không chậm, thậm chí còn vượt trội hơn Apriori và Eclat trên một số tập dữ liệu), mà là tính đơn giản của cấu trúc. Về cơ bản, mọi công việc đều được thực hiện trong một hàm đệ quy đơn giản, có thể được viết với tương đối ít dòng mã. [7]

### Ý tưởng

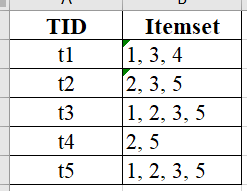
Lấy ý tưởng cơ bản của cả hai thuật toán FP\_Growth và H-mine, loại bỏ đệ quy dựa trên việc loại bỏ từng bước các mục khỏi cơ sở dữ liệu giao dịch cùng với quá trình xử lý đệ quy các tập hợp con giao dịch

### Thuật toán

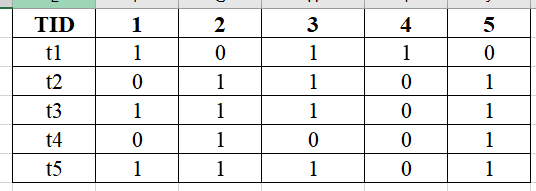
* Bước 1: Đếm tần suất các mục
* Bước 2: Loại bỏ các mục ít phổ biến
* Bước 3: Tái cấu trúc cơ sở dữ liệu
* Bước 4: Đệ quy loại bỏ và tìm kiếm
* Bước 5: Kết quả

### Ví dụ

Minsupp = 40% [6]



Hình 4.1: Ví dụ thuật toán RELim



Hình 4.2: CSDL sau khi biến đổi sang dạng nhị phân

* **Giai đoạn 1: các tập (1-itemsets)**

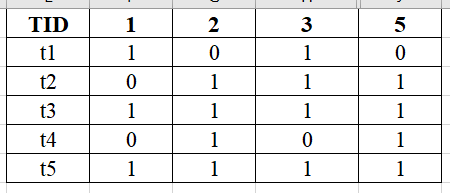
Ta có tập (1-itemsets) gồm {1}, {2}, {3}, {4}, {5}.

Tính độ hỗ trợ của những tập trên:

* supp({1}) = = 0.6 = 60% > minsupp (40%).
* supp({2}) = = 0.8 = 80% > minsupp (40%).
* supp({3}) = = 0.8 = 80% > minsupp (40%).
* supp({4}) = = 0.2 = 20% < minsupp (40%).
* supp({5}) = = 0.8 = 80% > minsupp (40%).
* Các tập {1}, {2}, {3}, {5} là tập phổ biến.
* Tập {4} không là tập phổ biến.

Vậy ta loại được tập {4}.

Giải thích: Dựa trên tính chất của tập phổ biến, nếu một tập không là tập phổ biến thì nhựng tập bao nó cũng không là tập phổ biến.



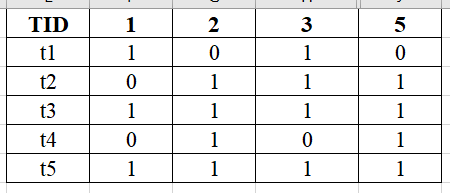
Hình 4.3: Dữ liệu sau giai đoạn 1

* **Giai đoạn 2: các tập (2-itemsets)**

Ta tiếp tục xây dựng các tập (2-itemsets) từ các tập {1}, {2}, {3}, {5}.

Tính độ hỗ trợ của những tập (2-itemsets):

* supp({1, 2}) = = 0.4 = 40% = minsupp (40%).
* supp({1, 3}) = = 0.6 = 60% > minsupp (40%).
* supp({1, 5}) = = 0.4 = 40% = minsupp (40%).
* supp({2, 3}) = = 0.6 = 60% > minsupp (40%).
* supp({2, 5}) = = 0.6 = 60% > minsupp (40%).
* supp({3, 5}) = = 0.6 = 60% > minsupp (40%).
* Các tập (2-itemsets) đều là tập phổ biến .
* Giai đoạn 2 không có tập nào bị loại bỏ.



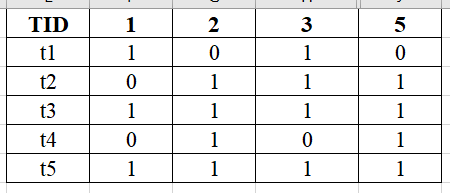
Hình 4.4: Dữ liệu sau giai đoạn 2

* **Giai đoạn 3: các tập (3-itemsets)**

Ta tiếp tục xây dựng các tập (3-itemsets) từ các tập {1}, {2}, {3}, {5}.

Tính độ hỗ trợ của những tập (3-itemsets):

* supp({1, 2, 3}) = = 0.4 = 40% = minsupp (40%).
* supp({1, 2, 5}) = = 0.4 = 40% = minsupp (40%).
* supp({1, 3, 5}) = = 0.4 = 40% = minsupp (40%).
* supp({2, 3, 5}) = = 0.6 = 60% > minsupp (40%).
* Các tập (3-itemsets) đều là tập phổ biến .
* Giai đoạn 3 không có tập nào bị loại bỏ.



Hình 4.4: Dữ liệu sau giai đoạn 3

* **Giai đoạn 4: các tập (4-itemsets)**

Ta tiếp tục xây dựng các tập (4-itemsets) từ các tập {1}, {2}, {3}, {5}.

Tính độ hỗ trợ của tập (4-itemsets): {1, 2, 3, 5}

* supp({1, 2, 3, 5}) = = 0.4 = 40% = minsupp (40%).
* Tập 1, 2, 3, 5} là tập phổ biến và là tập bao lớn nhất.

Kết luận: Ta có các tập phổ biến là {

{1}, {2}, {3}, {5},

{1, 2}, {1, 3}, {1, 5}, {2, 3}, {2, 5}, {3, 5},

{1, 2, 3}, {1, 2, 5}, {1, 3, 5}, {2, 3, 5}

{1, 2, 3, 5} }.

### Lưu ý

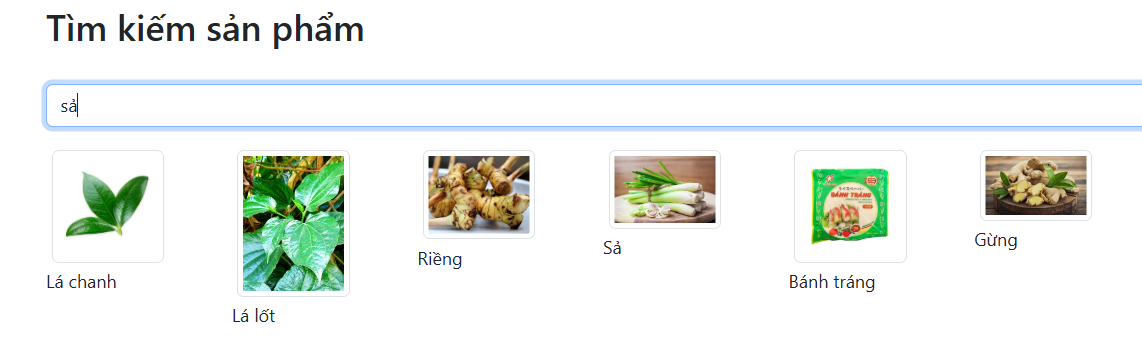
Khi tìm kiếm thuật toán RELim (Recursive Elimination) trên Google, rất dễ bị nhầm lẫn với thuật toán RFE (Recursive Feature Elimination) do tên của hai thuật toán khá giống nhau.

Có thể tìm kiếm bằng từ khóa “RELim Algorithm” hoặc “Recursive Elimination Algorithm” để có kết quả tốt hơn nhưng đa số sẽ là tài liệu bằng tiếng anh.

## Ứng dụng tìm tập phổ biến không theo thuật toán

### Giao diện

Tạo thành 1 web nhỏ demo về ứng dụng tập phổ biến vào thực tế, đề xuất sản phẩm mua kèm cho khách hàng



Hình 4.5: Giao diện ứng dụng

### Cách sử dụng

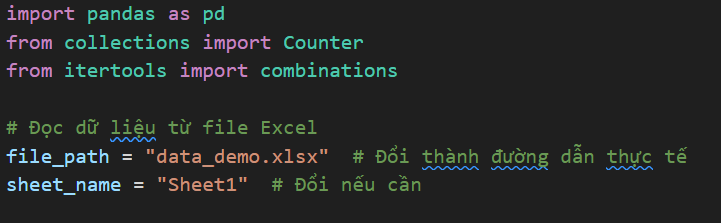
|  |  |
| --- | --- |
|  | * Data\_demo.xlsxlaf file cơ sỡ dữ liệu tổng hợp hoá đơn * Product\_images.json là file dùng để lưu trữ tên ảnh và đường đẫn ảnh * PhobienTD.py là file chính với mục đích là tìm kiếm ra được tập phổ biến tối đại * Domain.py là file dùng để tạo ra đường link web hoạt động như web server * Maximal\_frequent\_itemsets.json luôn trữ tập phổ biến dưới dạng từ điển |

Bước 1: Cài thư viện

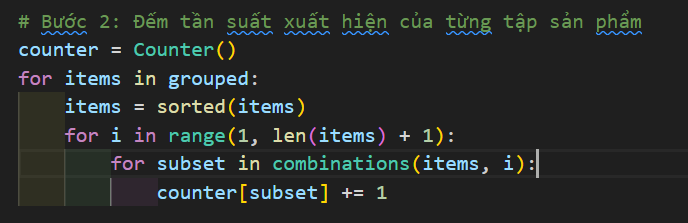
# pip install pandas ( Nếu chưa có)

Bước 2: Chạy file phobienTD.py để tìm ra tập phổ biến

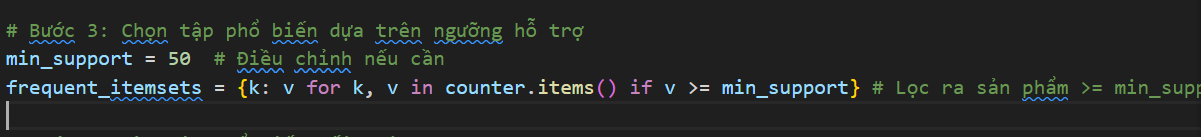
Bước 2.1: Thực hiện đọc dữ liệu từ file data.xlsx



Bước 2.2: Thực hiện đếm tần suất xuất hiện của từng sản phẩm có trong cơ sở dữ liệu

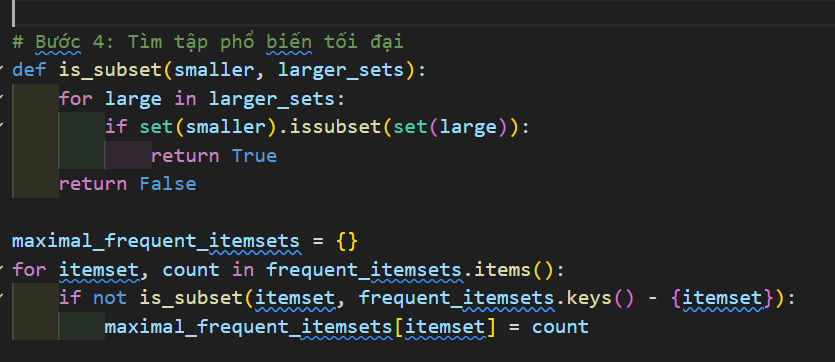


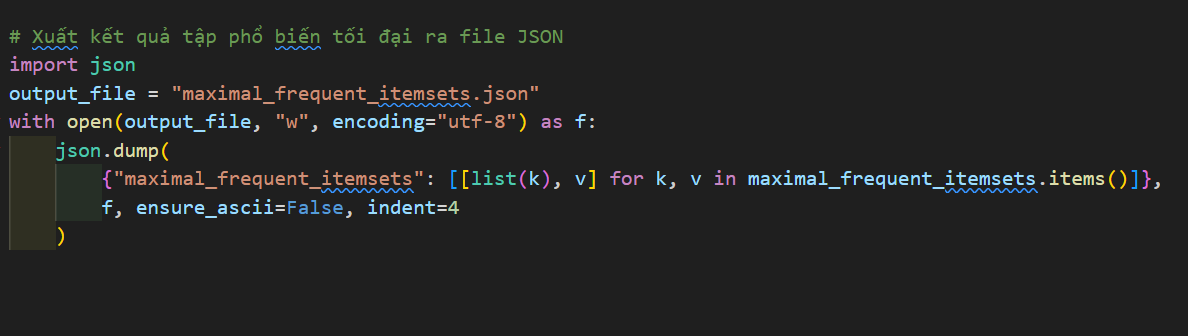
Bước 2.3: Chọn mức độ phổ biến min\_support để so sánh và lọc ra những sản phẩm là tập phổ biến



Bước 2.4: Tìm ra tập phổ biến tối đại, thông qua hàm is\_subset để tìm ra tất cả items đó nó có phải là tập con của large

Sau đó sẻ duyệt qua tất cả các items nằm trong tập phổ biến với điều kiện trừ cái itemset đang kiểm tra và not is\_subset( tức là tất cả các itemset không nằm trong 1 tập phổ biến nào lớn hơn nó)



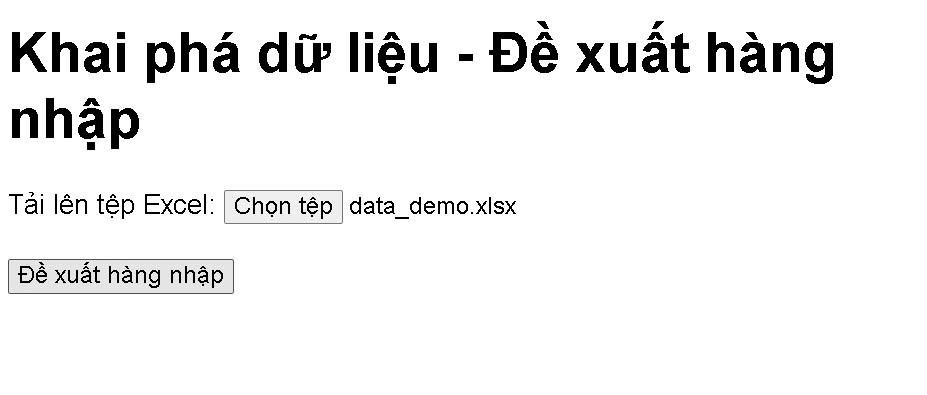
Bước 2.5: Xuất kết quả tập phổ biến tối đại thành file json

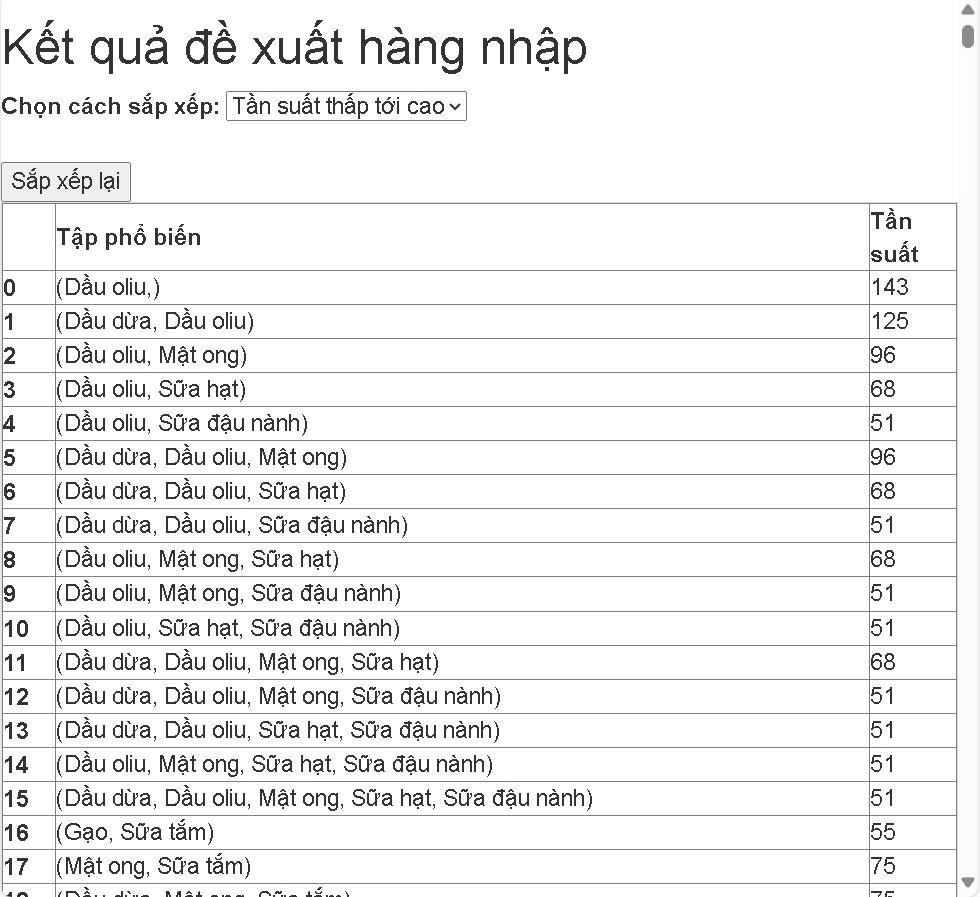
Bước 3: domain.py để tạo ra đường link tạo ra trang web

Bước 4: Thực hiện kiểm thử

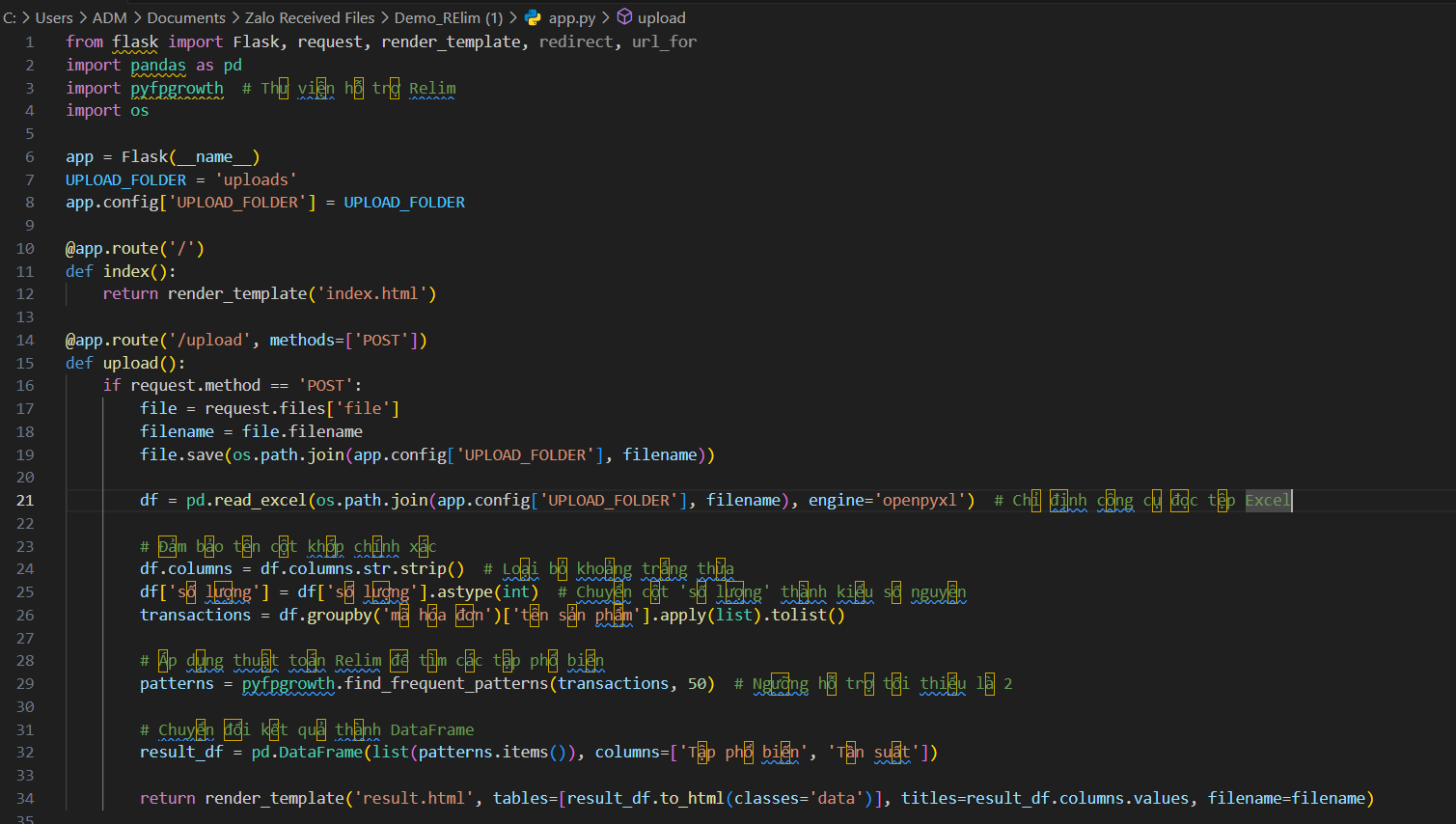
## Ứng dụng tìm tập phổ biến theo thuật toán Relim

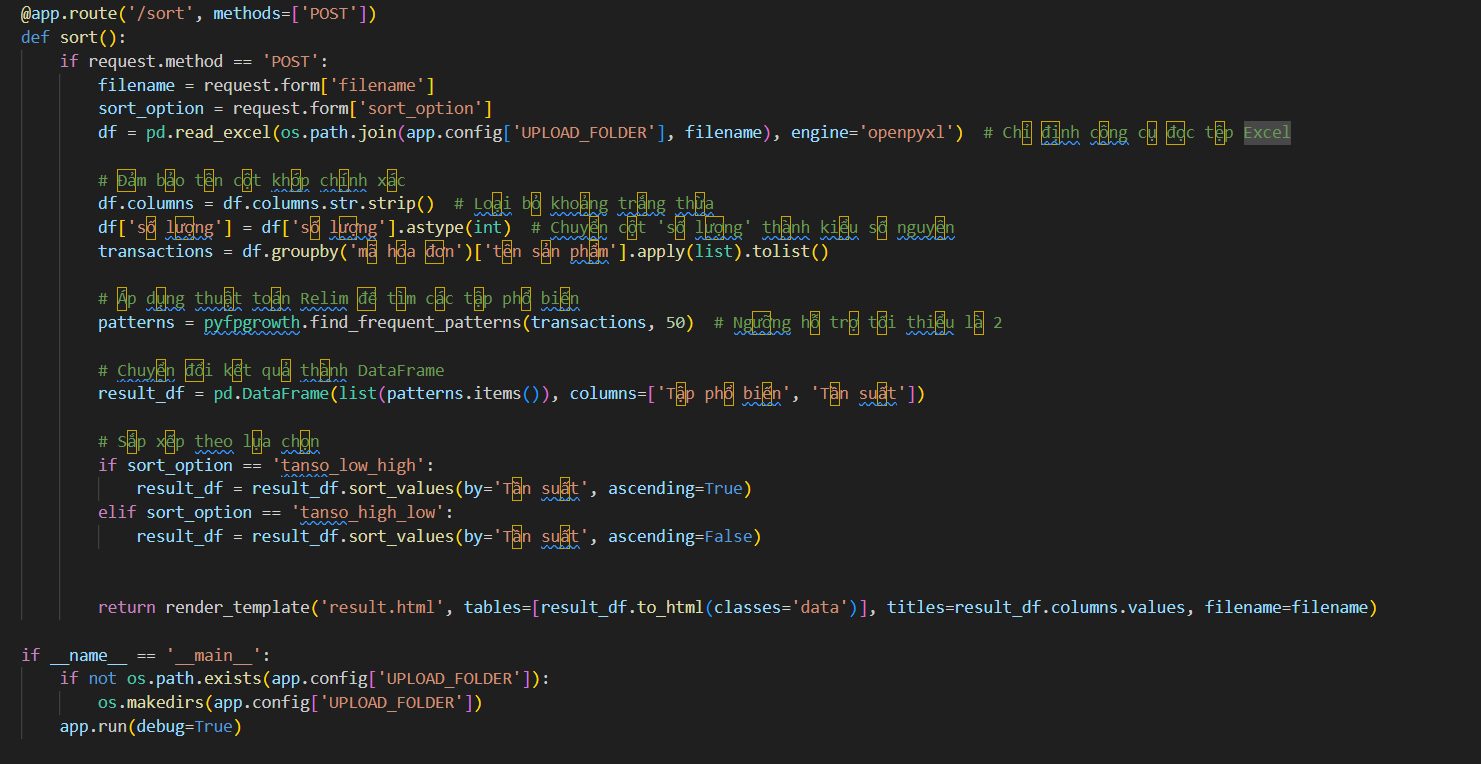
### Giao diện





### Code





TÀI LIỆU THAM KHẢO

[1] <https://www.youtube.com/watch?v=ErqF2l-iPcU&t=34s>

[2]<https://ongxuanhong.wordpress.com/2015/08/23/khai-thac-luat-tap-pho-bien-frequent-itemsets-voi-thuat-toan-apriori/>

[3]<https://viblo.asia/p/khai-pha-du-lieu-va-lop-bai-toan-khai-thac-cac-tap-pho-bien-djeZ1vX8lWz>

[4]<file:///C:/Users/ADM/Desktop/2767-V%C4%83n%20ba%CC%89n%20cu%CC%89a%20ba%CC%80i%20ba%CC%81o-5757-1-10-20210419.pdf>

[5] <file:///C:/Users/ADM/Desktop/tuannv,+08_Phan+Thanh+Huan_109-123.pdf>

[6][https://www.philippe-fournier-viger.com/spmf/RELim.php#:~:text=RELim%20is%20an%20algorithm%20for%20discovering%20itemsets%20(group%20of%20items,parameter%20given%20by%20the%20user](https://www.philippe-fournier-viger.com/spmf/Relim.php#:~:text=Relim%20is%20an%20algorithm%20for%20discovering%20itemsets%20(group%20of%20items,parameter%20given%20by%20the%20user).

[7][https://www.philippe-fournier-viger.com/spmf/RELim.pdf](https://www.philippe-fournier-viger.com/spmf/relim.pdf)

[8]<https://www.philippe-fournier-viger.com/spmf/index.php?link=documentation.php>