ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA HỆ THỐNG THÔNG TIN

BÁO CÁO ĐỒ ÁN

KHAI THÁC DỮ LIỆU

Đề tài: Recommend System

Giáo viên hướng dẫn: Nguyễn Hồ Duy Trí

Sinh viên thực hiện:

1. Lý Hoa Nam MSSV: 15520511
2. Đặng Xuân Phóng MSSV: 15520621

TP. Hồ Chí Minh, ngày 18 tháng 12 năm 2018

LỜI MỞ ĐẦU

Lời đầu tiên cho phép nhóm chúng em gửi lời cảm ơn chân thành và lòng biết ơn sâu sắc đến quý tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – ĐHQG TPHCM và quý Thầy Cô khoa Hệ thống thông tin, đặc biệt là ThS. Trịnh Minh Tuấn (Giảng viên dạy lý thuyết) đã truyền dạy những kiến thức cơ bản làm nên tảng để thực hiện đề tài và ThS. Nguyễn Hồ Duy Trí (giảng viên dạy thực hành) đã trực tiếp hướng dẫn, tận tình sửa chữa, đóng góp nhiều ý kiến, kinh nghiệm quý báu cho nhóm chúng em hoàn thành tốt cáo cáo môn học của mình.

Trong suốt quá trình thực hiện đề tài, nhóm chúng em đã vận dụng tối đa những kiến thức nền tảng đã tích luỹ đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới liên quan đến môn học Khai thác dữ liệu cũng như liên quan đến đề tài đang thực hiện để có thể hoàn thành đồ án một cách tốt nhất. Tuy nhiên, nhóm cũng gặp nhiều khó khăn nhất định nên những hạn chế, thiếu sót là điều khó tránh khỏi. Chính vì vậy nhóm chúng em rất mong nhận được sự góp ý, nhận xét từ phía Thầy để bổ sung kiến thức không chỉ để hoàn thiện cho đồ án lần này mà còn là hành trang cho những đồ án trong những học kì sắp tới.

Nhóm chúng em xin chân thành cảm ơn quý Thầy.

NHẬN XÉT CỦA GIÁO VIÊN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

MỤC LỤC

I. TỔNG QUAN ĐỀ TÀI.

1. Tổng quan về hệ thống gợi ý

Hệ thống gợi ý (Recommender systems hoặc Recommendation systems) là một dạng của hệ hỗ trợ ra quyết định, cung cấp giải pháp mang tính cá nhân hóa mà không phải trải qua quá trình tìm kiếm phức tạp. Hệ gợi ý học từ người dùng và gợi ý các sản phẩm tốt nhất trong số các sản phẩm phù hợp.

Hệ thống gợi ý sử dụng các tri thức về sản phẩm, các tri thức của chuyên gia hay tri thức khai phá học được từ hành vi con người dùng để đưa ra các gợi ý về sản phẩm mà họ thích trong hàng ngàn hàng vạn sản phẩm có trong hệ thống. Các website thương mại điện tử, ví dụ như sách, phim, nhạc, báo...sử dụng hệ thống gợi ý để cung cấp các thông tin giúp cho người sử dụng quyết định sẽ lựa chọn sản phẩm nào. Các sản phẩm được gợi ý dựa trên số lượng sản phẩm đó đã được bán, dựa trên các thông tin cá nhân của người sử dụng, dựa trên sự phân tích hành vi mua hàng trước đó của người sử dụng để đưa ra các dự đoán về hành vi mua hàng trong tương lai của chính khách hàng đó. Các dạng gợi ý bao gồm: gợi ý các sản phẩm tới người tiêu dùng, các thông tin sản phẩm mang tính cá nhân hóa, tổng kết các ý kiến cộng đồng, và cung cấp các chia sẻ, các phê bình, đánh giá mang tính cộng đồng liên quan tới yêu cầu, mục đích của người sử dụng đó.

2. Hệ thống gợi ý trong bài toán thực tiện

Để bán được sản phẩm nhiều hơn, giúp tăng doanh số cửa hàng A trên website. Cửa hàng A sự dụng hệ thống gợi ý khi khách hàng khi mua các sản phẩm. Cửa hàng A đã sử dụng dự liệu thu thập được khi khách hàng mua hàng trong một tháng. Để áp dụng thuận toán khai phá dữ liệu vào website.

2.1 Lý do chọn đề tài

Có thể nói rằng hiện nay có đủ mọi loại thông tin trên internet và với sự giúp sức của các công cụ tìm kiếm thông tin hiện hữu, người dùng có thể tìm thấy những gì họ quan tâm. Mặc khác, do có quá nhiều thông tin, nên gây ra nhiều khó khăn, lúng túng cho việc ra quyết định lựa chọn của người dùng. Các hệ khuyến nghị (recommendation systems) ra đời không nằm ngoài mục đích hỗ trợ cho người dùng trong các lựa chọn ra quyết định.

Có nhiều hướng tiếp cận để xây dựng một hệ khuyến nghị. Tùy thuộc vào nguồn thông tin có được, nhu cầu khuyến nghị thực tế, đặc thù riêng của dịch vụ khuyến nghị cần cung cấp,… mà mỗi hệ khuyến nghị sẽ có phương pháp và thuật toán phù hợp cho riêng mình. Kỹ thuật (Frequent Itemsets) là một trong số các kỹ thuật được sử dụng để xây dựng một hệ khuyến nghị dựa trên các dữ liệu đánh giá. Kỹ thuật này được đánh giá cao nhờ khả năng cải thiện độ chính xác của các thuật toán khuyến nghị khác, tính linh hoạt, thời gian thực thi thấp, ...

2.2 Mục tiêu đề tài

Việc chọn đề tài “Nghiên cứu phương pháp (Frequent Itemsets) và ứng dụng vào hệ khuyến nghị sản phẩm” nhằm mục tiêu nâng cao kết quả khuyến nghị thông tin sản phẩm cho người dùng. Bằng cách kết hợp với phân tích quan điểm người dùng đối với hạng mục. Chúng tôi mong rằng hệ thống sẽ khuyến nghị các hạng mục phù hợp với nhu cầu của người dùng. Từ đó hỗ trợ cho người dùng trong các lựa chọn ra quyết định.

2.3 Phạm vi nghiên cứu

Vì thời gian thực hiện đề tài có giới hạn, chúng tôi chỉ tập trung nghiên cứu những nội dung sau:

- Tìm hiểu tổng quan hệ khuyến nghị.
- Tìm hiểu về thuật toán fpgrowth và apiory.
- Tìm hiểu công cụ và ngôn ngữ xây dựng ứng dụng minh họa

2.4 Công nghệ và công cụ sử dụng

- Phía server:
  - Flask, Pandas, Apiory, Fpgrowth : là Micro Framework dành cho Python
  - NodeJS: một mã nguồn được xây dựng dựa trên nền tảng Javascript V8 Engine.
  - Một số thư viện hổ trợ khác.
- Phía client:
  - ReactJs thư viện front-end chạy trên nền tẳng Javascript
  - Bootstrap, Thư viện CSS, Javascript hỗ trợ xây dựng giao diện.
  - Và một số thư viện hỗ trợ khác.

II. MÔI TRƯỜNG PHÁT TRIỂN

1. Python 3

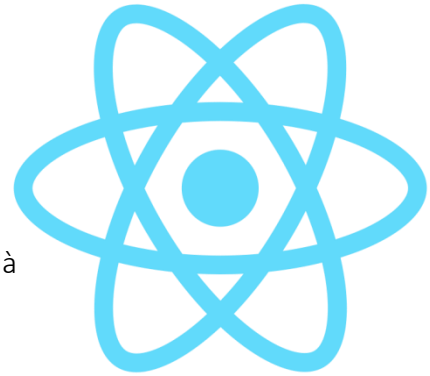
**Sử dụng python 3 để khởi chạy thuật toán*

Python là một ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng, do Guido van Rossum tạo ra và lần đầu ra mắt vào năm 1991. Python được thiết kế với ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ. Python là ngôn ngữ có hình thức rất sáng sủa, cấu trúc rõ ràng, thuận tiện cho người mới học lập trình. Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu. Vào tháng 7 năm 2018, Van Rossum đã từ chức Leader trong cộng đồng ngôn ngữ Python sau 30 năm lãnh đạo

2. ReactJs

React.js là 1 thư viện JavaScript tạo ra bởi Facebook.

(http://facebook.github.io/react/)



Như khái niệm trên trang web chính thức "A JavaScript library for building user interface", React.js là một thư viện sinh ra để xây dựng giao diện người dùng (UI). Nó không phải là Framework mà chỉ là thư viện, do đó trong MVC nó sẽ tương ứng với phần V.

III. XÂY DỰNG MÔ HÌNH KHUYẾN NGHỊ

1 Mô hình tổng quan

Dựa trên mô hình tổng quan, các công việc cần được thực hiện như sau:

- Dữ liệu được có sẵn và lưu trong file csv.
- Lấy dữ liệu và chạy các thuật toán khai phá
- Hiện thị thuật toán trên giao diện người dùng

2. Dataset

2.1. Mô tả dataset

- Dataset được sử dụng trong bài là Dataset "store_data.csv".
- Dataset "store_data.csv" được pulished trên trang web: https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/
- Dataset mô tả sản phẩm được mua của một cửa hàng tại Pháp.
- Mục tiêu: Để giới thiệu sản phẩm (B,C,..) được mua tương tự khi mua sản phẩm(A)
- Thông số Dataset:

+ Số dòng: 7501
+ Số thuộc tính: 240
+ Dung lượng: 296KB
+ Thuộc tính phân lớp: Y

shrimp	almonds	avocado	vegetable	green gra	whole we	yams	cottage ch	energy dri	tomato ju	low fat yo	green tea	honey	salad	mineral w	salmon	antioxyda	frozen sm	spinach	olive oil
burgers	meatballs	eggs																	
chutney																			
turkey	avocado																		
mineral w	milk	energy ba	whole wh	green tea															
low fat yogurt																			
whole wh	french fries																		
soup	light crear	shallot																	
frozen ve	spaghetti	green tea																	
french fries																			
eggs	pet food																		
cookies																			
turkey	burgers	mineral w	eggs	cooking oil															
spaghetti	champagn	cookies																	
mineral w	salmon																		
mineral water																			
shrimp	chocolate	chicken	honey	oil	cooking oi	low fat yogurt													
turkey	eggs																		
turkey	fresh tuna	tomatoes	spaghetti	mineral w	black tea	salmon	eggs	chicken	extra dark	chocolate									
meatballs	milk	honey	french frie	protein bar															
red wine	shrimp	pasta	pepper	eggs	chocolate	shampoo													
rice	sparkling	water																	
spaghetti	mineral w	ham	body spra	pancakes	green tea														
burgers	grated che	shrimp	pasta	avocado	honey	white win	toothpaste												
eggs																			
parmesan	spaghetti	soup	avocado	milk	fresh bread														
ground be	spaghetti	mineral w	milk	energy ba	black tea	salmon	frozen sm	escalope											
sparkling water																			
mineral w	eggs	chicken	chocolate	french fries															
frozen ve	spaghetti	yams	mineral water																
herb & pe	tomato sa	light crear	magazines																
mineral w	chocolate	avocado	eggs																
turkey	french frie	strawberries																	
frozen ve	strong che	chocolate																	

2.2. Bước tiền xử lý

Sử dụng ngôn ngữ Python để tiền xử lý dữ liệu.

```
import pandas as pd
store_data = pd.read_csv("store_data.csv",header=None, keep_default_na=False)
```

Dùng thư viện pandas để đọc file csv.

```
records = []
#dataset have col: 20, row: 7501
for i in range(0,7501):
    temps = []
    for j in range(0,20):
        if (dataset.values[i,j] == ''):
            break
        else:
            temps.append(str(dataset.values[i,j]))
    records.append(temps)
```

-Chuyển dữ liệu từ dạng dataFrame sang dạng mạng và bỏ các phần tử NaN để chuyển sang dạng ma trận:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	shrimp	almonds	avocado	vegetables mix	green grapes	whole weat flour	yams	cottage cheese	energy drink	tomato juice	low fat yogurt	green tea	honey	salad	mineral water	salmon	antioxydant juice	frozen smoothie
1	burgers	meatballs	eggs	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	chutney	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	turkey	avocado	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	mineral water	milk	energy bar	whole wheat rice	green tea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

3. Khai phá luật kết hợp

3.1Thuật toán apriori

Sử dụng ngôn ngữ Python và thư viện Apiory (*https://github.com/ymoch/apyori*)

```
associantion_rules = apriori(records, min_support=0.0045, min_confidence=0.2, min_li
association_results=list(associantion_rules)
```

Kết quả :

3.2. Thuật toán FP-Growth

Sử dụng ngôn ngữ python và thư viện fp-growth (https://github.com/evandempsey/fp-growth)

```
    patterns = fp.find_frequent_patterns(records, minsup)
    rules = list(fp.generate_association_rules(patterns, minconf))
```

kết quả

3.3 Đưa dữ liệu lên API

Chuyển dữ sang dạng Json và format lại dữ liệu

```
    data = {}
    arr= []
    for item in association_results:

        values = {}
        pair = item[0]
        items = [x for x in pair]
        values["fist"] = [items[0]]
        values["next"] = [items[1]]

        arr.append(values)
    data["rules"]= arr
    data["min_sup"]= minsup
    data["min_conf"]=minconf
    return data
```

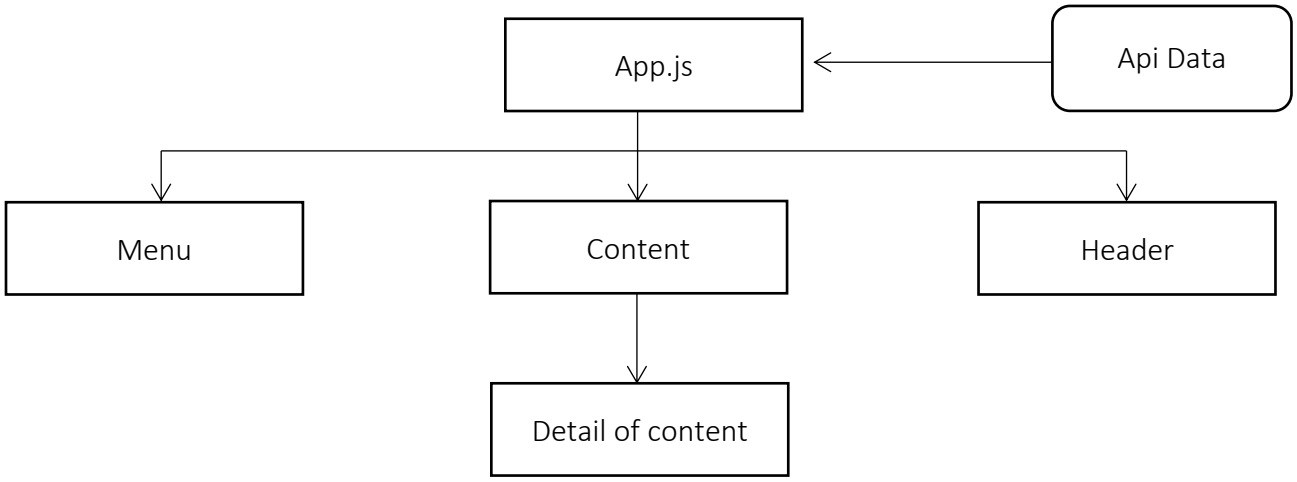
Kết quả khi dùng thư viện trả về với dạng list, với các format khó sử dụng. Phần code trên format lại kết quả phù hợp với bài toán.

```python
app = flask.Flask(__name__)
cors = CORS(app, resources={r"/api/*": {"origins": "*"}})
app.config["DEBUG"] = True
@app.route('/', methods=['GET'])
def home():
    return '''<h1>Distant Reading Archive</h1> '''
@app.route('/api/fpgrowth')
def api_fp():
    return jsonify(result_fpgrowth)
@app.route('/api/apiori')
def api_ap():
    return jsonify(result_apyori)
app.run()
```

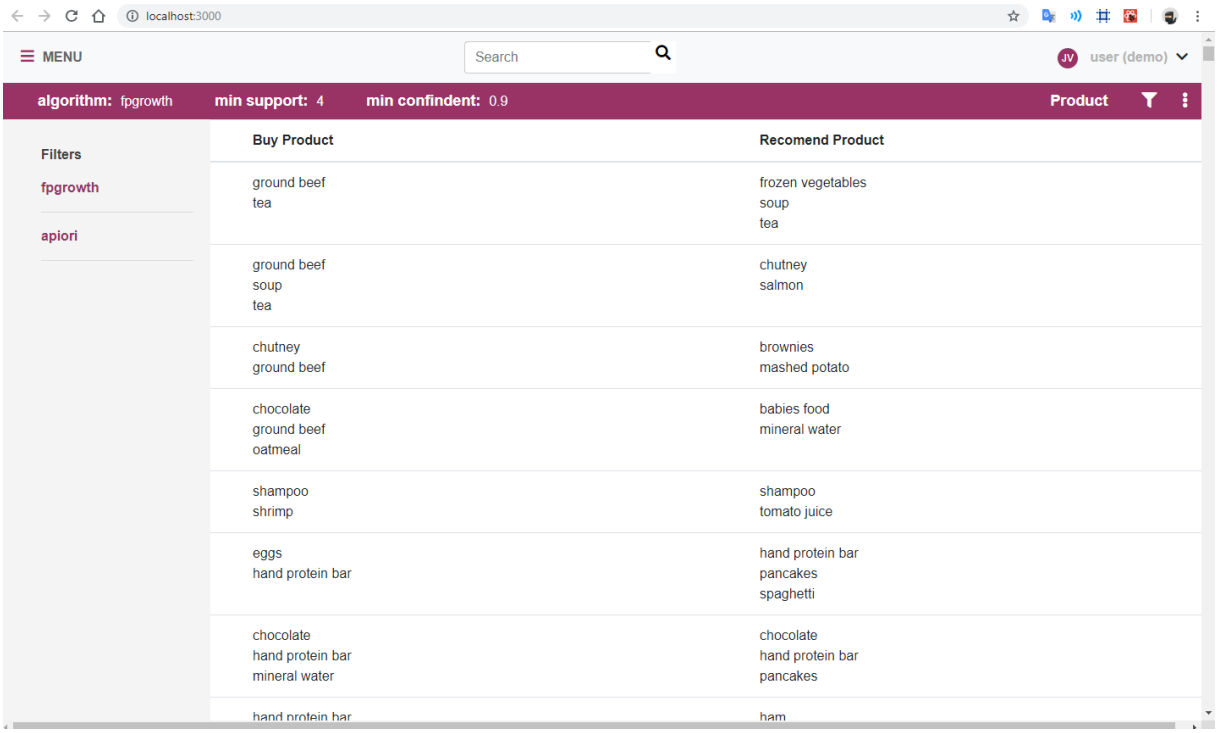
Sử dụng thư viện flask để tạo server với port 5000 (http://127.0.0.1:5000/api/...) và trả file json lên server.

4. Hiện thị dữ liệu

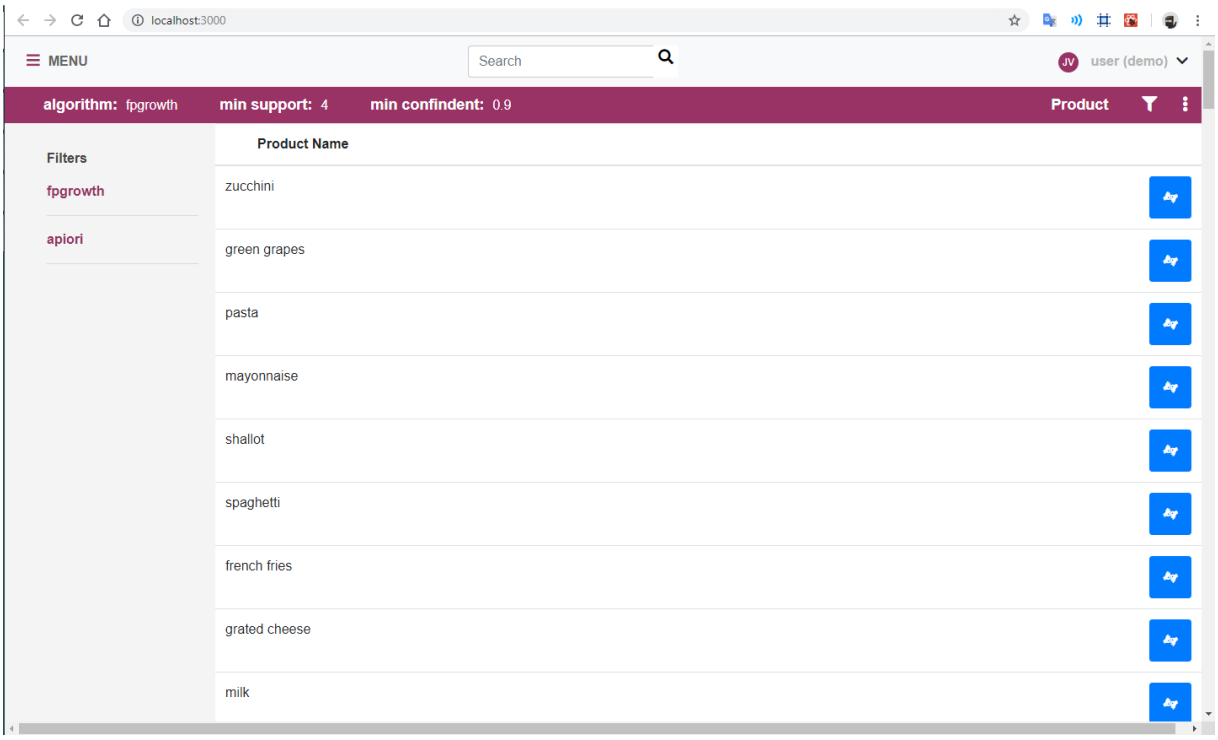
4.1 Mô hình hiện thị



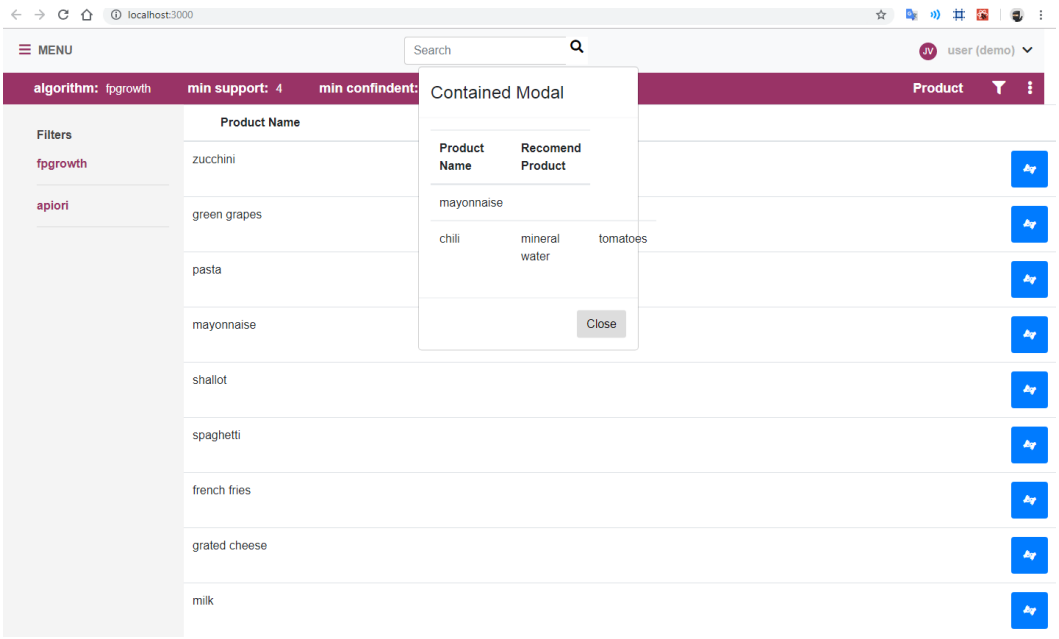
4.2 Giao diện người dùng



Hiện thị thuật toán



Hiện thị sản phẩm.



Đưa ra luật khi chọn sản phẩm mayonase

IV. KẾT LUẬN

1. Đánh giá kết quả:

Độ chính xác:

-Do thuật toán apyori được hỗ trợ thêm phần lift và length còn thuật toán fpgrowth ko có. Nên khó đánh giá chính xác về thuật toán.

Tốc độ chạy:

-> Thuật toán fpgrowth có tốc độ chạy nhanh hơn, dựa trên khi chạy thực nghiệm cùng minsup, mincof và được đánh giá trên các diễn đàn github

So sánh thư viện thuật toán sử dụng

->Thư viện fpgrowth được hỗ trợ tốt hơn về phần input và output của dữ liệu. Nhưng lại không hỗ trợ về độ chính xác tốt như thư viên apyori

2. Ưu và nhược điểm của ứng dụng.

Ưu điểm

Sau thời gian nghiên cứu và thực hiện, chúng tôi nhận thấy hệ thống cũng đạt được một số ưu điểm sau:

- Ứng dụng đáp ứng được nhu cầu hiện thị kết quả thuật toán.
- Có các chức năng tìm kiếm
- UI/UX tương đối dễ sử dụng
- Tốc độ sử lý tương đối tốt.

Nhược điểm

Trong quá trình thực hiện đề tài, nhóm tác giả không thể tránh khỏi những sai sót. Sau đây, một số khuyết điểm mà chúng tôi nhận thấy:

- Độ chính xác chưa thật sự được cao

- Chưa đánh giá được kết quả khi chạy thuật toán
- Còn nhiều chức năng chưa hoàn thành

3. Hướng phát triển

Như đã trình bày, hệ thống hiện tại vẫn còn một số hạn chế, chúng tôi nhận định rằng vẫn còn nhiều việc phải làm để cải tiến và hoàn thiện hệ thống. Vì vậy, chúng tôi đề ra các mục tiêu trong tương lai như sau:

- Chạy nhiều thuật toán recommend system hơn.
- Import được nhiều dataset khác.
- Chức năng tìm kiếm thông minh hơn
- Có thêm chức năng filter sản phẩm
- Thiết kế giao diện tối ưu hơn

V. PHÂN CÔNG

1. Phân công

1. Đặng Xuân Phóng

-Thực hiện tìm thư viện

-Chạy thư viện trên môi trường python

-Tiền xử lý dữ liệu

2. Lý Hoa Nam

-Chuyển kết quả về format phù hợp

-Đưa dữ liệu lên API

-Hiện thị dữ liệu dựa vào thư viện reactjs

2. Phiếu đánh giá

Tên	Điểm	Ghi chú
Lý Hoa Nam	**8**	
Đặng Xuân Phóng	**8**	

V. Tài liệu tham khảo

1. A Python implementation of the Frequent Pattern Growth algorithm https://pypi.org/project/pyfpgrowth/
2. Association Rule Mining via Apriori Algorithm in Python https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/
3. Association Rules Generation from Frequent Itemsets http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/#association-rules-generation-from-frequent-itemsets
4. Creating Web APIs with Python and Flask (Patrick Smyth) https://programminghistorian.org/en/lessons/creating-apis-with-python-and-flask
5. Build simple Medium.com on Node.js and React.js https://codeburst.io/build-simple-medium-com-on-node-js-and-react-js-a278c5192f47

6. Creating a Simple Recommender System in Python using Pandas https://stackabuse.com/creating-a-simple-recommender-system-in-python-using-pandas/
7. Beginner's Tutorial on the Pandas Python Library https://stackabuse.com/beginners-tutorial-on-the-pandas-python-library/
8. React documentation https://reactjs.org/docs/getting-started.html
9. Python 3.7.2rc1 documentation https://docs.python.org/3/
10. Silde và bài giảng.