

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM **TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG KHOA CÔNG NGHỆ THÔNG TIN**

**BÁO CÁO CUỐI KÌ**

**MÔN: XỬ LÝ DỮ LIỆU LỚN**

**TÌM HIỂU VỀ FREQUENT ITEMSET**

*Người hướng dẫn*: **GV LÊ ANH CƯỜNG**

*Người thực hiện*: **NGUYỄN ĐẠI HIỆP – 51900683**

**NGUYỄN MINH TRÍ – 51900251**

**NGUYỄN PHƯỚC VINH - 51900283**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**



TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM **TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG KHOA CÔNG NGHỆ THÔNG TIN**

**BÁO CÁO CUỐI KÌ**

**MÔN: XỬ LÝ DỮ LIỆU LỚN**

**TÌM HIỂU VỀ FREQUENT ITEMSET**

*Người hướng dẫn*: **GV LÊ ANH CƯỜNG**

*Người thực hiện*: **NGUYỄN ĐẠI HIỆP – 51900683**

**NGUYỄN MINH TRÍ – 51900251**

**NGUYỄN PHƯỚC VINH - 51900283**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn quý thầy cô và Trường Đại học Tôn Đức Thắng đã tạo điều kiện và cơ hội cho chúng em biết đến và học hỏi được rất nhiều kiến thức bổ ích về môn Xử lý dữ liệu lớn. Trong suốt quá trình học tập môn Xử lý dữ liệu lớn, chúng em đã có thể nhận được những kiến thức bổ ích qua sự tận tình từ những bài học trên giảng đường của thầy, và cả sự đóng góp tích cực của các bạn học đã giúp chúng em có một quãng thời gian trải qua môn học thật sự ý nghĩa và vui vẻ. Những nội dung mà thầy đã giảng dạy và chia sẻ cho chúng em trong quá trình học tập đã giúp ích cho chúng em rất nhiều như: sự gắn bó đoàn kết giữa các thành viên để hoàn thành các bài kiểm tra, làm báo cáo giữa kì cuối kì, và đặc biệt là chúng em học được rất nhiều kiến thức hữu ích để áp dụng vào thực tế cho quá trình trưởng thành và làm việc sau này.

Đặc biệt, nhóm chúng em xin chân thành cảm ơn thầy Lê Anh Cường là người đã truyền đạt trực tiếp lại cho chúng em rất nhiều kiến thức và kinh nghiệm, nhiệt tình hướng dẫn và giúp đỡ, dẫn lối cho chúng em trong suốt quá trình học tập, mặc dù tụi em chỉ có cơ hôi học trực tuyến vì tình hình dịch bệnh Covid 19 diễn biến phức tạp, cũng như đã cho tụi em sự tự tin, kĩ năng và kiến thức để hoàn thành bài báo cáo giữa kì lần này. Trong quá trình làm bài báo cáo, chúng em sẽ không tránh khỏi những sai sót, mong thầy có thể nhận xét và góp ý để nhóm chúng em có thể nhận ra lỗi sai, những lỗ hổng kiến thức và đặc biệt là sự khắc phục để có thể áp dụng kiến thức cho quá trình học hỏi được những kinh nghiệm mới và có thể hoàn thiện bản thân cho những bài báo cáo tiếp theo và cả quá trình làm việc sau này.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Chúng tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi và được sự hướng dẫn của Thầy Lê Anh Cường;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Nguyễn Đại Hiệp*

*Nguyễn Minh Trí*

*Nguyễn Phước Vinh*

# TÓM TẮT

Thế giới đang phát triển với một tốc độ nhanh chóng, tất cả các ngàng nghề đều cũng phát triển theo. Chung quy lại, những gì con người đã và đang nghiên cứu tất cả cùng có một xu hướng là phục vụ đời sống con người. Lĩnh vực công nghệ thông tin cũng không ngoại lệ, các hệ thống trang web, ứng dụng, trí tuệ nhân tạo, … đang có những sản phẩm rất chất lượng phục vụ cho con người và lĩnh vực đang rất phát triển hứa hẹn nhiều điều trong tương lai.

Nhận thấy điều đó, khoa Công nghệ thông tin thuộc trường Đại học Tôn Đức Thắng đã đưa ra các đề tài nghiên cứu về lĩnh vực xử lý dữ liệu lớn – một nhánh nhỏ của công nghệ thông tin,nhằm mục đích giúp sinh viên hiểu rõ hơn về cách thức hoạt động cũng như hiệu quả mà chúng mang lại cho con người. Cụ thể trong nhóm đề tài ấy, chúng em chọn đề tài Frequent Itemset để nghiên cứu và sử dụng thuật toàn Apriori trong chương trình minh họa cho đề tài của chúng em.

Hiện tại chúng em đã tìm hiểu và phân tích về đề tài Frequent Itemset và viết được một chương trình python sử dụng thuật toán Apriori để giải quyết bài toán. Chương trình hoạt động khá tốt và đưa ra những kết quả gần đúng nhất. Và điều quan trong nhất là trong quá trình nghiên cứu, chúng em thực hiện chương trình trong môi trường Pyspark thì thu được kết quả nhanh hơn rất nhiều so với thông thường. Cụ thể về quá trình nghiên cứu và kết quả chi tiết thì xin mọi người cùng xem các phần tiếp theo của báo cáo.

# MỤC LỤC

[TÓM TẮT 5](#_Toc90993053)

[MỤC LỤC 6](#_Toc90993054)

[CHƯƠNG 1 – GIỚI THIỆU 7](#_Toc90993055)

[1.1. Giới thiệu: 7](#_Toc90993056)

[1.2. Mục đích: 7](#_Toc90993057)

[1.3. Phương pháp giải quyết bài toán: 7](#_Toc90993058)

[***1.3.1. Apriori 7***](#_Toc90993059)

[***1.3.2. Frequent Pattern Growth 8***](#_Toc90993060)

[1.4. Phương pháp nghiên cứu: 8](#_Toc90993061)

[*1.4.1. Support 8*](#_Toc90993062)

[*1.4.2. Confident 9*](#_Toc90993063)

[*1.4.3. Lift 9*](#_Toc90993064)

[CHƯƠNG 2 – CHƯƠNG TRÌNH THỂ HIỆN THUẬT TOÁN 10](#_Toc90993065)

[2.1. Thư viện 10](#_Toc90993066)

[2.2. Không song song 11](#_Toc90993067)

[2.3. Song song 13](#_Toc90993068)

[2.4. Đo thời gian của hàm 13](#_Toc90993069)

[2.5. Đồ thị đo thời gian của hàm 14](#_Toc90993070)

[KẾT LUẬN 15](#_Toc90993071)

[TÀI LIỆU THAM KHẢO 16](#_Toc90993072)

# CHƯƠNG 1 – GIỚI THIỆU

## Giới thiệu:

* Bài toán khai thác tập phổ biến (frequent itemset) là bài toán rất quan trong trong lĩnh vực khai phá dữ liệu. Bài toán khai thác các tập phổ biến được ứng dụng trong rất nhiều vấn đề, nổi tiếng nhất là Basket data analysis (dự đoán, gợi ý các món hàng thường được cho cùng vào giỏ hàng với item đã được chọn mua trước đó). Ngoài ra, nó còn được ứng dụng trong các bài toán: phân tích thua lỗ, phân cụm, phân loại, hệ thống khuyến nghị, … Và đặc biệt nó còn thể ứng dụng cho thiết kế các dịch vụ tiện ích trong smarthome.

## Mục đích:

* Tìm tất cả các tập mẫu, liên kết, tương quan hoặc cấu trúc nhân quả có độ lặp cao trong tập hợp tất cả các hạng mục hoặc đối tượng trong cơ sở dữ liệu giao dịch, cơ sở dữ liệu quan hệ và các kho thông tin dữ liệu khác.

## Phương pháp giải quyết bài toán:

* Frequent itemset có thể khai thác thông qua 2 thuật toán phổ biến:
* Apriori
* FP Growth

1. *Apriori*

* Apriori là một trong những thuật toán đầu tiên và là nền móng phát triển cho các thuật toán khai thác các tập phổ biến sau này.
* Thuật toán Apriori được công bố bởi R. Agrawal và R. Srikant vào năm 1994 vì để tìm các tập phổ biến trong một bộ dữ liệu lớn. Tên của thuật toán là Apriori vì nó sử dụng kiến thức đã có từ trước (prior) về các thuộc tính, vật phẩm thường xuyên xuất hiện trong cơ sở dữ liệu. Để cải thiện hiệu quả của việc lọc các mục thường xuyên theo cấp độ, một thuộc tính quan trọng được sử dụng gọi là thuộc tính Apriori giúp giảm phạm vi tìm kiếm của thuật toán.
* Việc thuật toán Apriori có thể làm là nhìn vào quá khứ và khẳng định rằng nếu một việc gì đó xảy ra thì sẽ có tỉ lệ bao nhiêu phần trăm sự việc tiếp theo sẽ xảy ra. Nó giống như nhìn vào quá khứ để dự đoán tương lại vậy, và việc này rất có ít cho các nhà kinh doanh. Ví dụ một siêu thị muốn nghĩ cách sắp xếp các gian hàng một cách hợp lí nhất, họ có thể nhìn vào lịch sử mua hàng và sắp sếp các tập sản phẩm thường được mua cùng nhau vào một chỗ. Hoặc một trang web tin tức muốn giới thiệu cho người dùng các bài viết liên quan đến nhau nhất, cũng có thể áp dụng quy luật tương tự.
* Nhược điểm của thuật toán Apriori là thời gian thực hiện rất lớn lãng phí nhiều thời gian ở quá trình sinh ra cá tập ứng cử viên candidate itemsets. Không gian tìm kiếm lớn và chi phí tính toán cũng rất cao.

1. *Frequent Pattern Growth*

* Khám phá các tập phổ biến frequent itemsets mà không cần sinh ra các tập ứng cử viên candidate itemsets. Ưu điểm của nó là xây dựng cơ sở dữ mẫu điều kiện từ cơ sở dữ liệu thỏa mãn độ hỡ trợ tối thiểu. Do cấu trúc cây nhỏ gọn và không tạo ra các tập ứng cử viên candidate itemsets nên yêu cầu ít bộ nhớ hơn. Nhược điểm của nó là hoạt động kém với các tập dữ liệu có mẫu dài (độ dài k của itemset lớn).

## Phương pháp nghiên cứu:

* Từ những ưu điểm mà Apriori mang lại, nhóm chúng em lựa chọn thuật toán Aprori.
* Thuật toán Apriori bao gồm 3 thành phần chính:
* Confident
* Support
* Lift

1. *Support*

* Support đề cập đến mức độ phổ biến của item trong itemset. Ta có thể tính support bằng cách lấy số giao dịch chứ item chia cho tổng số giao dịch.

support (item1) = (Số giao dịch(item1))/tổng số giao dịch

* Ví dụ ta có một bài bài toán cụ thể như sau: có tập dữ liệu A chứa n lần khách hàng mua sản phẩm công nghệ, trong đó có 100 khách hàng mua iphone6.

support (iphone6) = 100/n

1. *Confident*

* Confident là tỉ lệ mà item1 được mua cùng với item2 nếu mặt hàng item2 được mua. Và confident của item1 khi mua item2 được tính bằng cách lấy tổng số lần item1 và item2 được mua cùng nhau chia cho tổng số lần item2 được mua.

confident (item2→item1) = số lần giao dịch (item1 & item2) / số lần giao dịch (item2)

* Cũng lấy tập dataset A của ví dụ trên làm minh họa, có 100 lần mua cả item1 và item2 và 500 lần chỉ mua item2:

confident (item2→item1) = 100/500 = 20%

1. *Lift*

* Lift (item2→item1) là sự gia tăng tỉ lệ item1 được bán khi item2 được bán.
* Lift(item2→item1) được tính bằng cách lấy confident (item2→item1) chia cho support (item1).

lift (item2→item1) = confident (item2→item1) / support (item1)

* Lấy ví dụ tập dataset A có tổng 10000 lượt 100 lượt giao dịch cùng mua item1 và item2, 200 lượt chỉ mua item1, 500 lượt giao dịch chỉ mua item2.

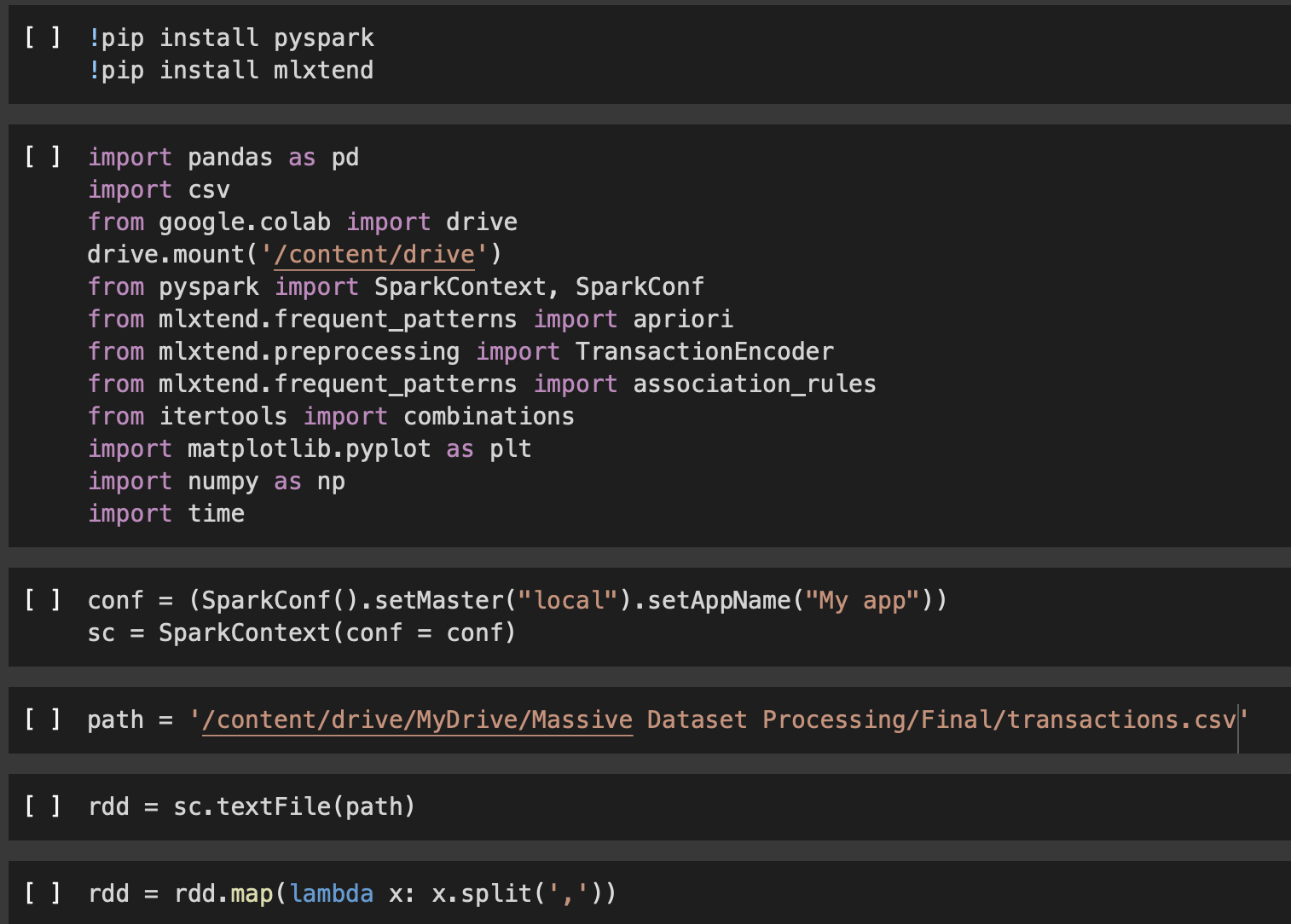
confident (item1→item2) = 100/500 = 20%=0.2

support (item1) = 200/10000 = 2% = 0.02

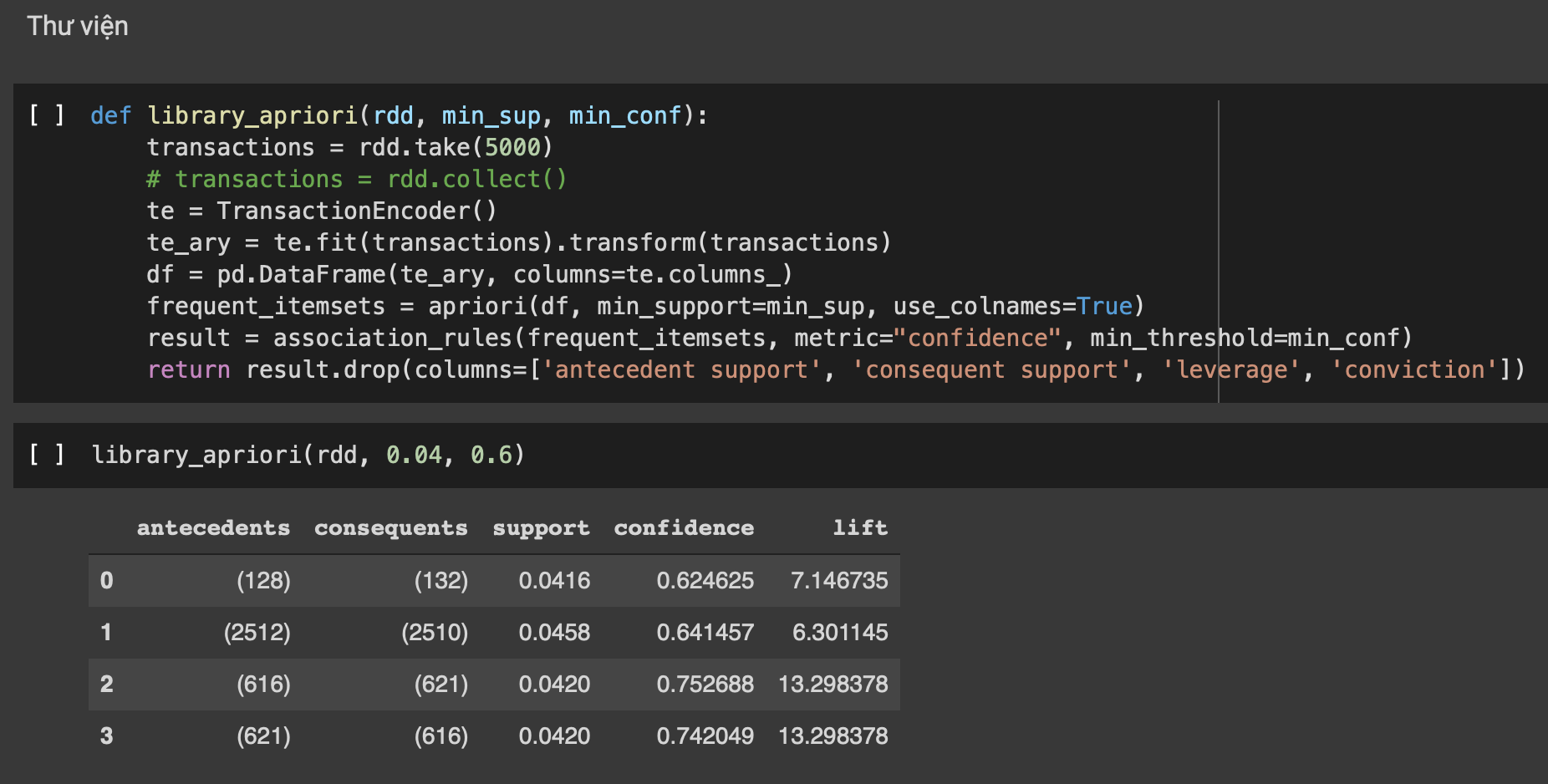
lift (item2→item1) = confident (item2→item1) / support (item1) = 0.2/0.02 = 10

* Vậy từ ví dụ trên có thể kết luậ rằng tỉ lệ mua một lượt cả item1 và item2 cao hơn 10 lần so với chỉ mua item1.

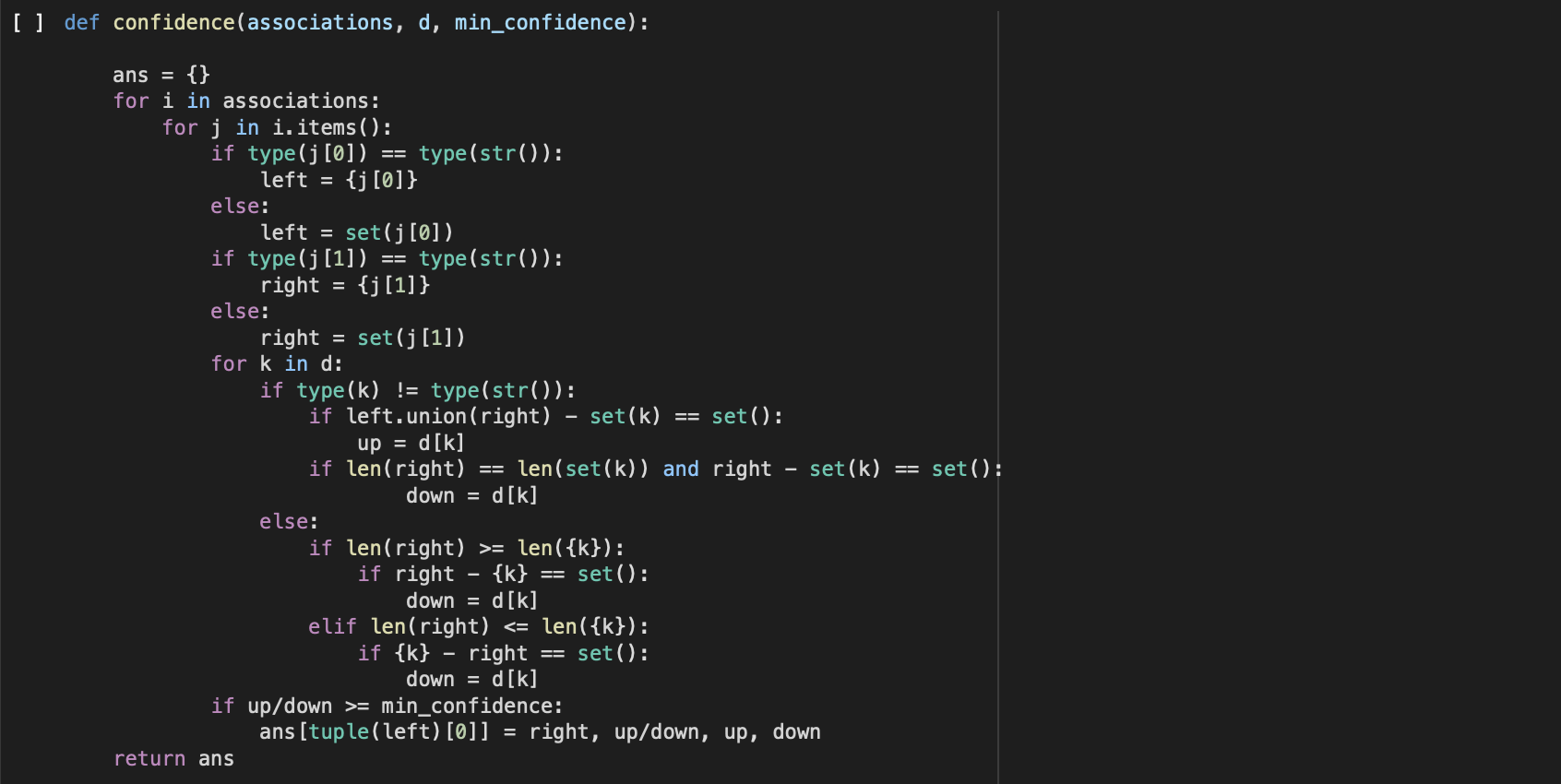
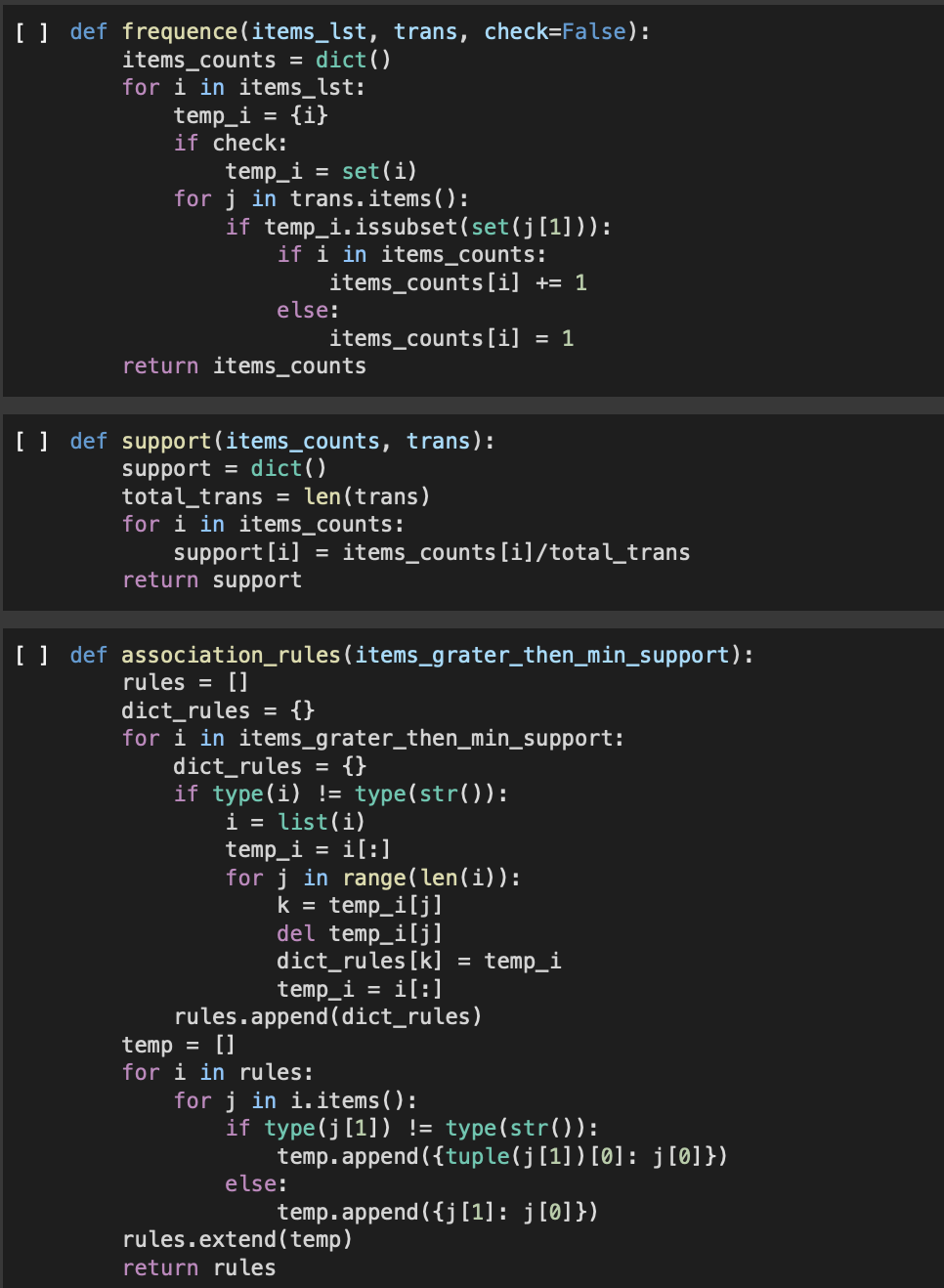
# CHƯƠNG 2 – CHƯƠNG TRÌNH THỂ HIỆN THUẬT TOÁN

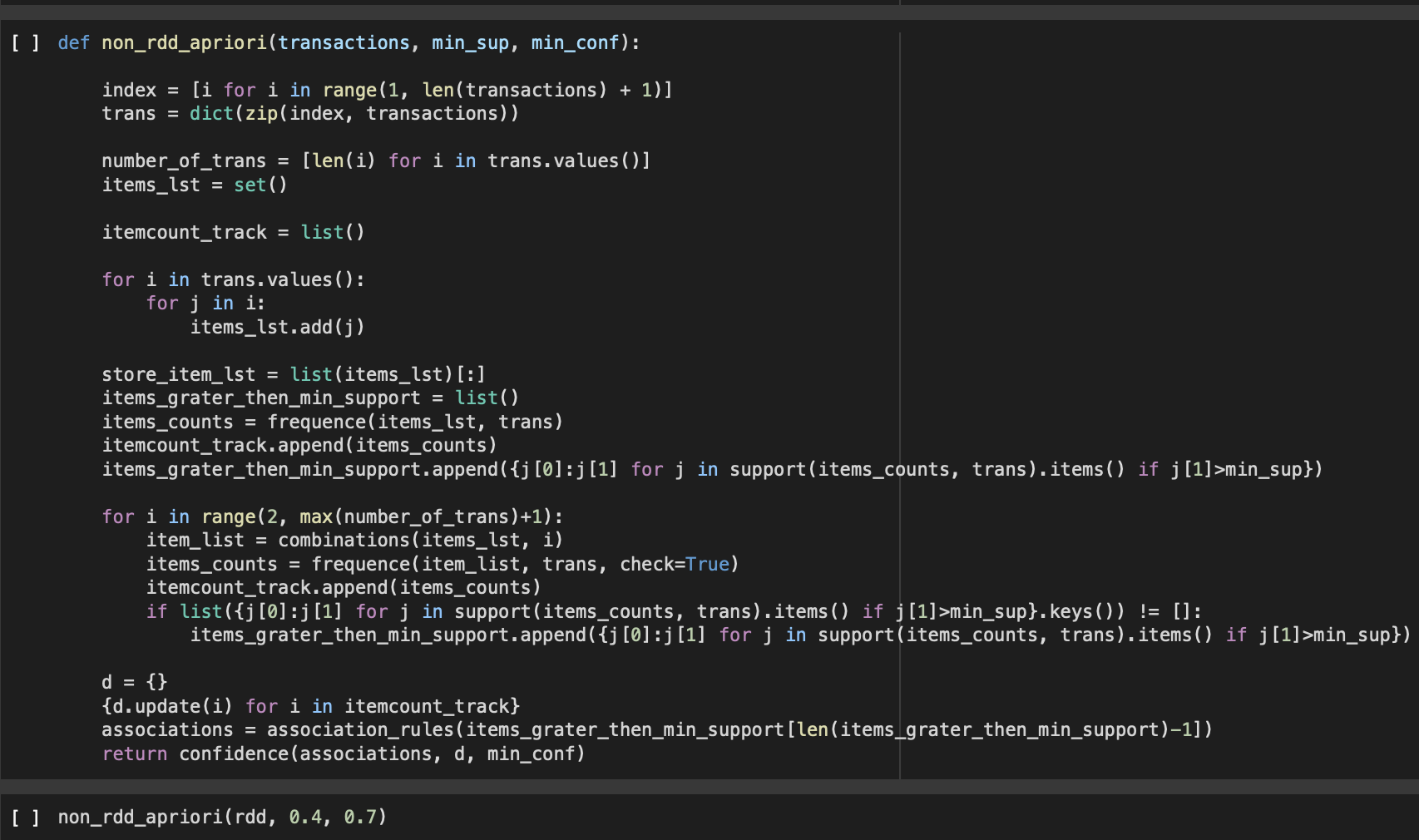


1. **Thư viện**

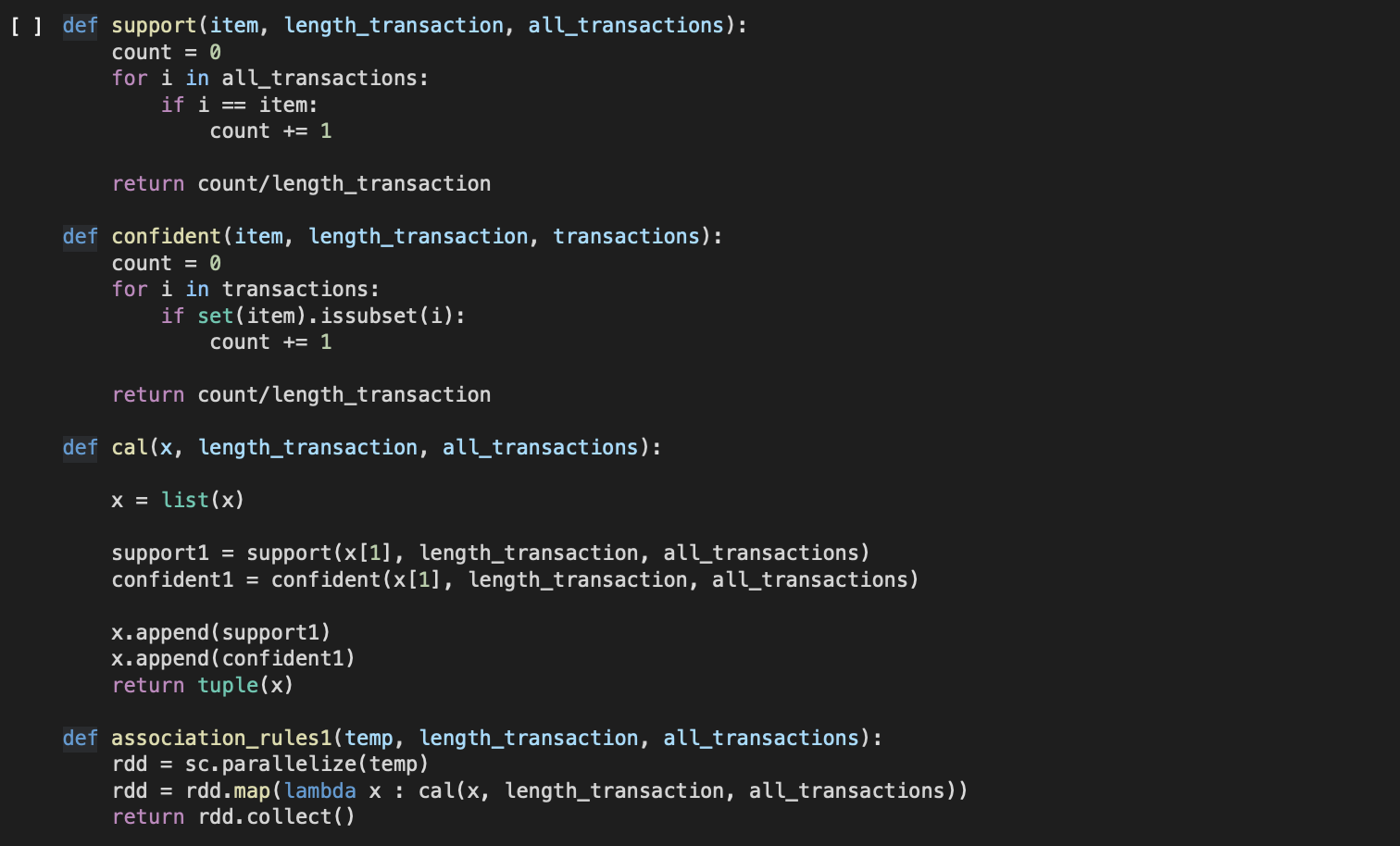


1. **Không song song**



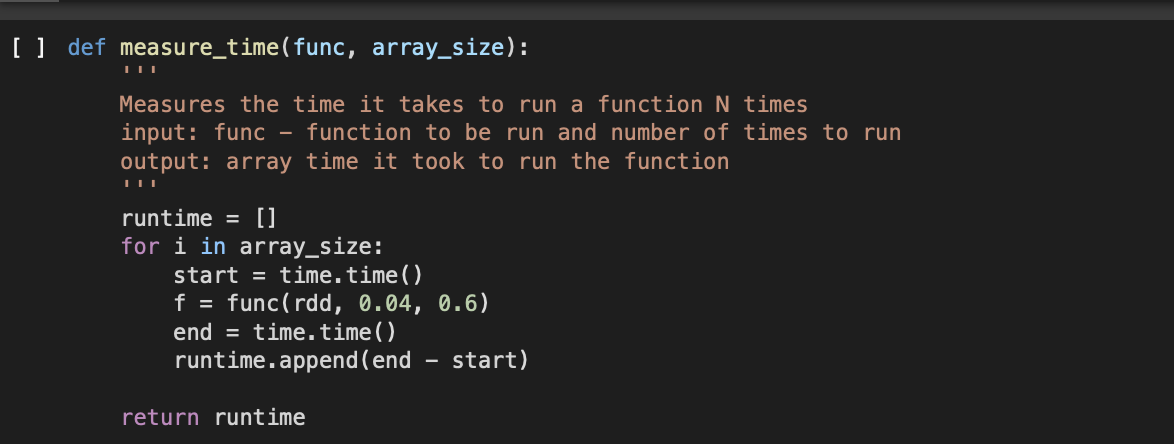


1. **Song song**

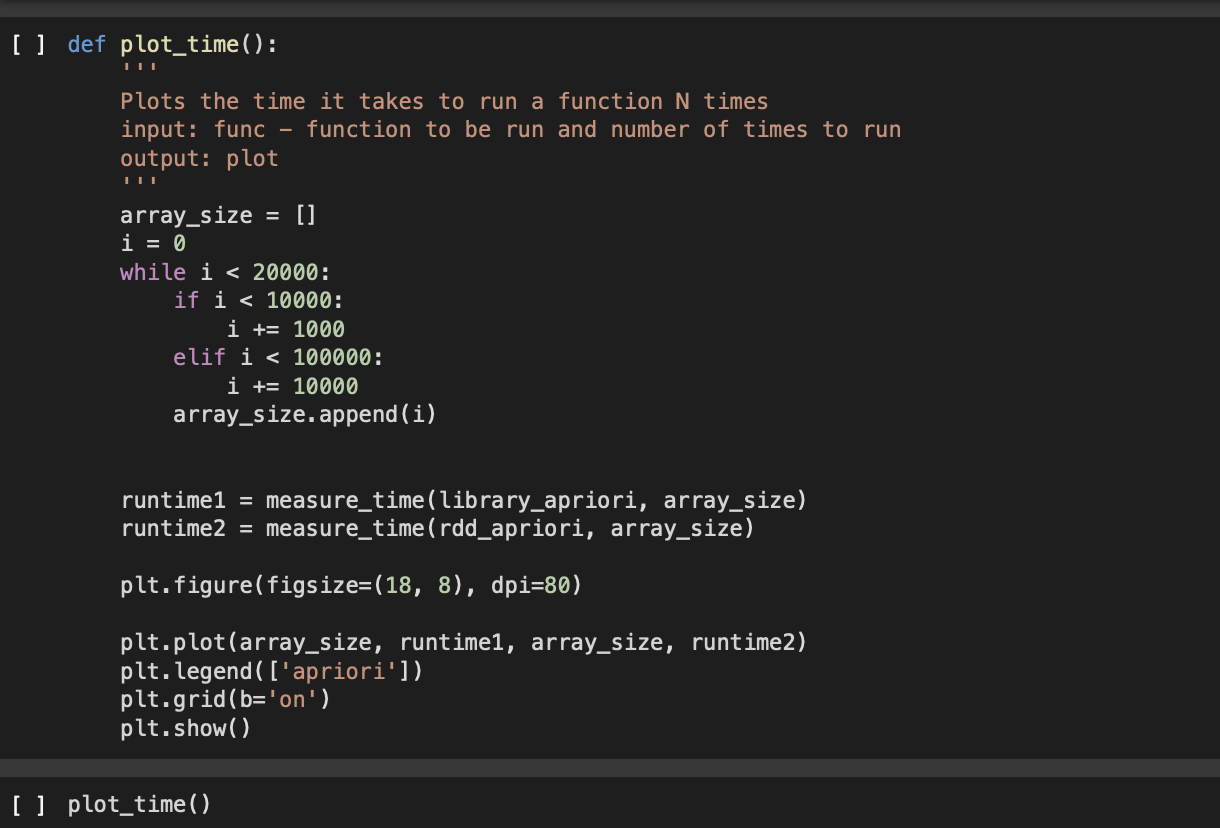
****

****

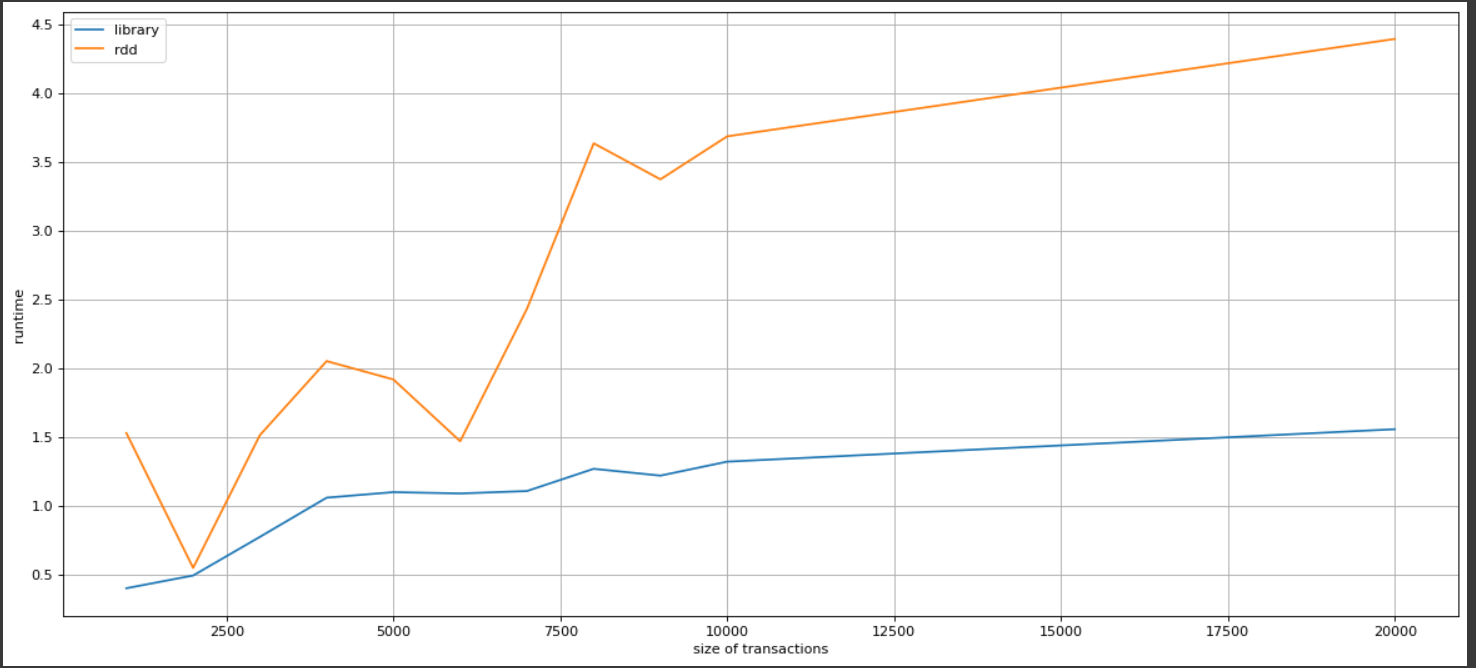
1. **Đo thời gian của hàm**

****

1. **Đồ thị đo thời gian của hàm**

****

# KẾT LUẬN



*Chú thích:*

Đường vàng: thời gian chạy khi sử dụng thuật toán song song

Đường xanh: thời gian chạy khi sử dụng thư viện Apriori

# TÀI LIỆU THAM KHẢO

1. <https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/>
2. <https://www.geeksforgeeks.org/implementing-apriori-algorithm-in-python/>
3. <http://www.lirmm.fr/~lazaar/imagina/NL-DM-IMAGINA1819-part1.pdf>
4. <https://viblo.asia/p/khai-pha-du-lieu-va-lop-bai-toan-khai-thac-cac-tap-pho-bien-djeZ1vX8lWz>
5. <http://www.ungdung.khoa-hnvd.com/Hoc_thuat/Apriori.html>
6. <https://ichi.pro/vi/hieu-va-xay-dung-thuat-toan-fp-growth-bang-python-238291391263554>
7. <https://www.geeksforgeeks.org/ml-frequent-pattern-growth-algorithm/>