GVHD: LÊ NGỌC THÀNH, NGUYỄN NGỌC THẢO

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
THÀNH PHỐ HỒ CHÍ MINH

BÀI TẬP 2: KHAI THÁC LUẬT KẾT HỢP

I. THÔNG TIN SINH VIÊN

Họ và tên: **TRẦN NHẬT HUY**

Mssv: **1612272**

Email: nhathuy13598@gmail.com

Sđt: **0354 878 677**

II. BẢNG BÁO CÁO CÔNG VIỆC

STT	CÁC CÂU HỎI	MỨC ĐỘ HOÀN THÀNH	GHI CHÚ
A	1. Tìm hiểu về phương pháp cải tiến quá trình tìm luật kết hợp từ tập phổ biến	100%	
	2.a. Sử dụng thuật toán Apriori và FP-Growth. Liệt kê tập phổ biến tối đại và tập phổ biến đóng	100%	
	2.b. Tìm tất cả các luật kết hợp	100%	
	2.c. Ứng dụng cải tiến ở câu a. vào câu b.	100%	
B	1. Chuyển dữ liệu trong plants.data sang dạng nhị phân	100%	
	2. Trả lời các câu hỏi	100%	
	3. Áp dụng thuật toán Apriori	100%	
	4. Khai thác tập phổ biến	100%	
	5. Khai thác luật kết hợp	100%	

III. CHI TIẾT BÀI LÀM

A. LÝ THUYẾT

1. Tìm hiểu về một phương pháp cải tiến khai thác luật kết hợp từ tập phổ biến

Nhược điểm của phương pháp khai thác luật kết hợp dựa trên tập phổ biến là khi số lượng tập phổ biến lớn thì luật kết hợp sinh ra sẽ rất lớn. Trong đó, một luật kết hợp này có thể là con của luật kết hợp khác do đó chúng ta sẽ phải tiến hành loại bỏ các luật là con của luật khác

Các phương pháp được đề nghị sử dụng thay thế cho tập phổ biến là khai thác luật kết hợp trên tập phổ biến đóng và tập phổ biến tối đại

Tập phổ biến đóng là tập phổ biến mà không có tập nào bao nó có cùng độ phổ biến

Tập phổ biến tối đại là tập phổ biến mà không có tập nào bao nó cũng là tập phổ biến

Từ định nghĩa ta có : tập phổ biến tối đại ⊆ tập phổ biến đóng ⊆ tập phổ biến tối đại

Tập phổ biến đóng thể hiện đầy đủ thông tin của tập phổ biến cùng với độ hỗ trợ chính xác của nó. Luật kết hợp được lấy ra từ tập phổ biến đóng sẽ nhỏ gọn hơn, dễ quản lý và phân tích. Giả sử, từ tập phổ biến ta tạo ra được tập phổ biến đóng, tuy nhiên tập phổ biến đóng này vẫn còn quá lớn thì khi đó ta sẽ tìm tập phổ biến tối đại. Khai thác tập phổ biến tối đại thích hợp với cơ sở dữ liệu dày đặc

Để khai thác tập phổ biến đóng, ta sử dụng thuật toán **CHARM**. Để khai thác tập phổ biến tối đại thì ta sử dụng thuật toán **GenMax**. Cả 2 thuật toán trên đều sử dụng cây được tạo ra từ thuật toán **Eclat**

2. Cài đặt các thuật toán

a. *Sử dụng thuật toán **Apriori** và **FP-Growth** để tìm luật kết hợp. So sánh. Tìm tập phổ biến tối đại và tập phổ biến đóng*

Thuật toán **Apriori** và **FP-Growth** được cài đặt trong các file lần lượt là **Apriori.py** và **FP_Growth.py**

Tập phổ biến khi chạy **Apriori.py** và **FP_Growth.py**

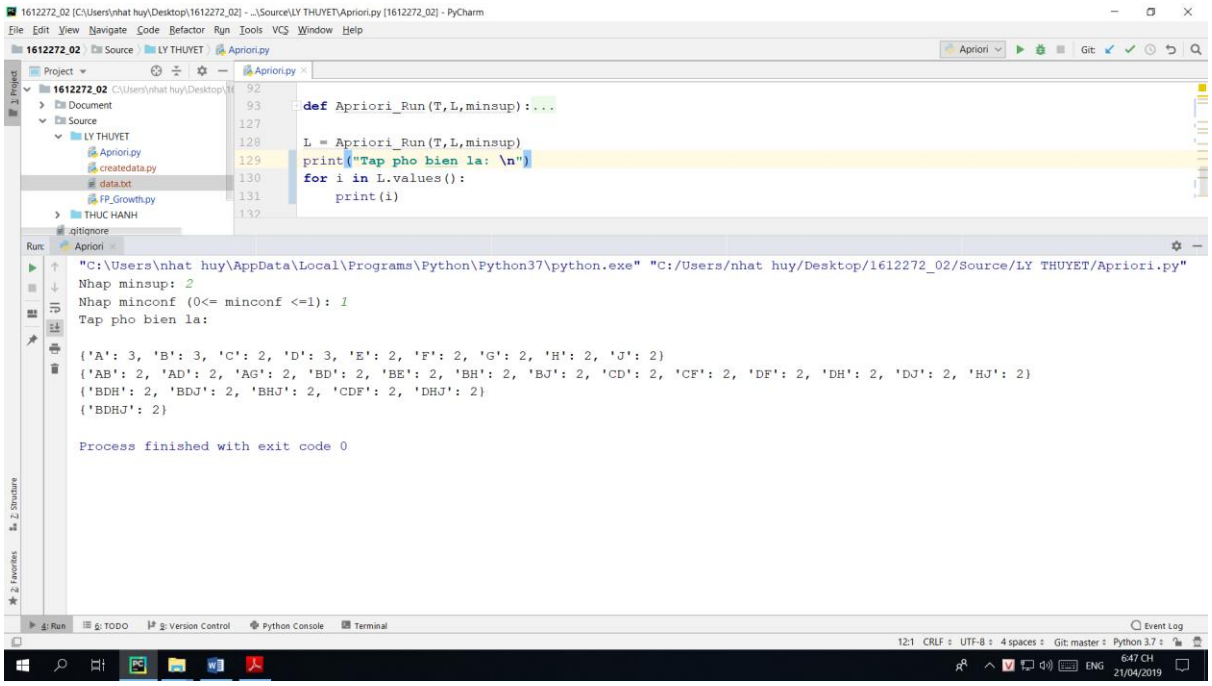


Figure 1: Tập phổ biến khi sử dụng Apriori.py

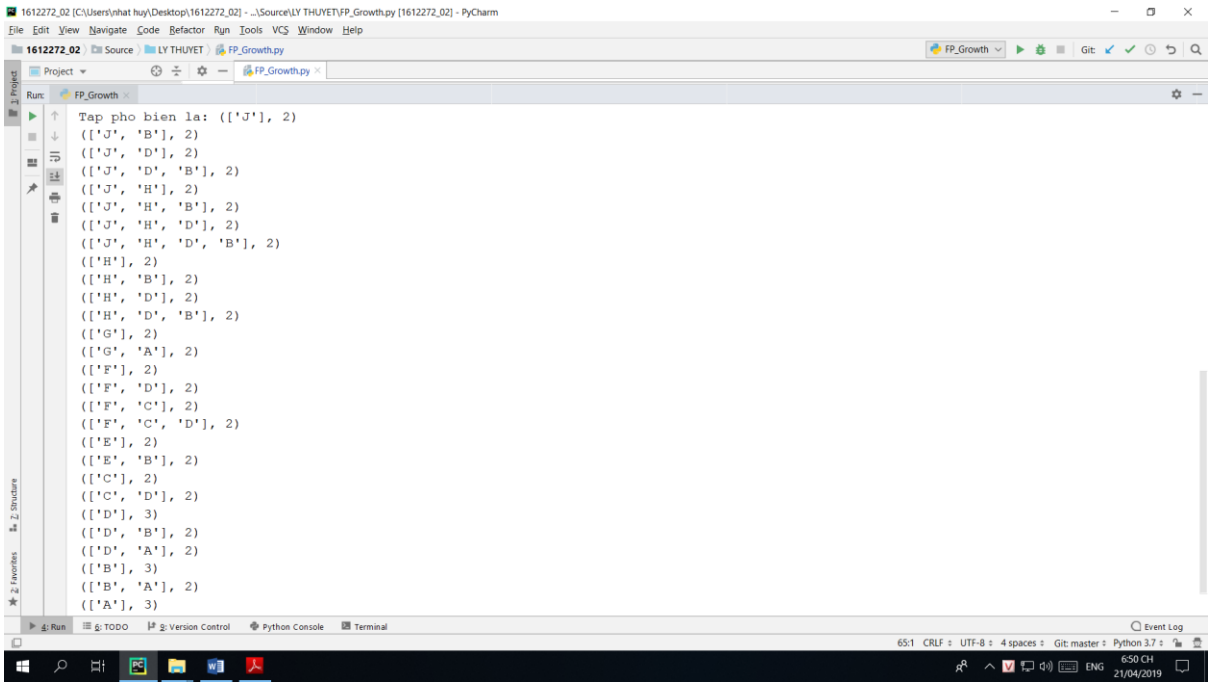


Figure 2: Tập phổ biến khi chạy FP_Growth.py

Kết quả của 2 phương pháp này là giống nhau

K-ITEM SET	ITEM
1-item set	A:3, B:3, C:2, D:3, E:2, F:2, G:2, H:2, J:2
2-item set	AB:2, AD:2, AG:2, BD:2, BE:2, BH:2, BJ:2, CD:2, CF:2, DF:2, DH:2, DJ:2, HJ:2
3-item set	BDH:2, BDJ:2, BHJ:2, CDF:2, DHJ:2
4-item set	BDHJ:2

Ta chạy file **Apriori.py** để tìm tập phổ biến tối đại và tập phổ biến đóng

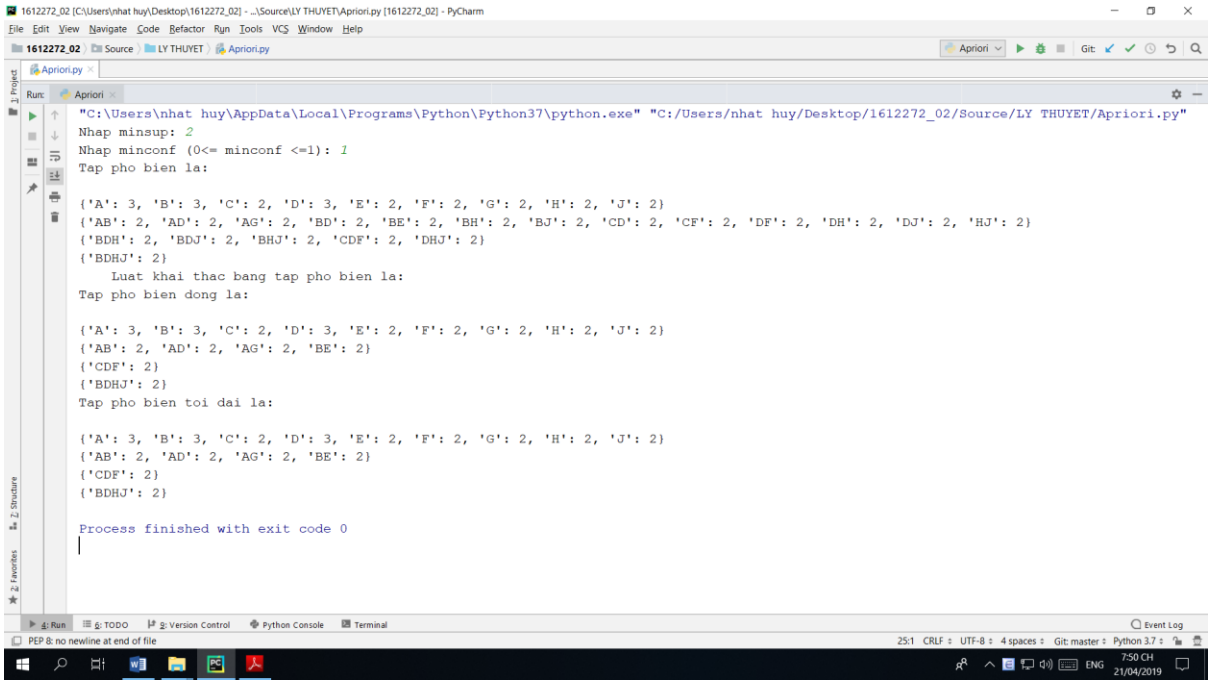


Figure 3: Tập phổ biến tối đại và tập phổ biến đóng

Ta giữ lại tập **1-item set** để dễ tính **minconf**

Tập phổ biến đóng

K-ITEM SET	ITEM
1-item set	A:3, B:3, C:2, D:3, E:2, F:2, G:2, H:2, J:2
2-item set	AB:2, AD:2, AG:2, BD:2, BE:2
3-item set	CDF:2
4-item set	BDHJ:2

Tập phổ biến tối đại

K-ITEM SET	ITEM
1-item set	A:3, B:3, C:2, D:3, E:2, F:2, G:2, H:2, J:2
2-item set	AB:2, AD:2, AG:2, BD:2, BE:2
3-item set	CDF:2
4-item set	BDHJ:2

b. *Tìm tất cả luật kết hợp thỏa* ***minsup*** *và* ***minconf***

Chạy file **Apriori.py** để tìm các luật kết hợp thỏa **minconf**

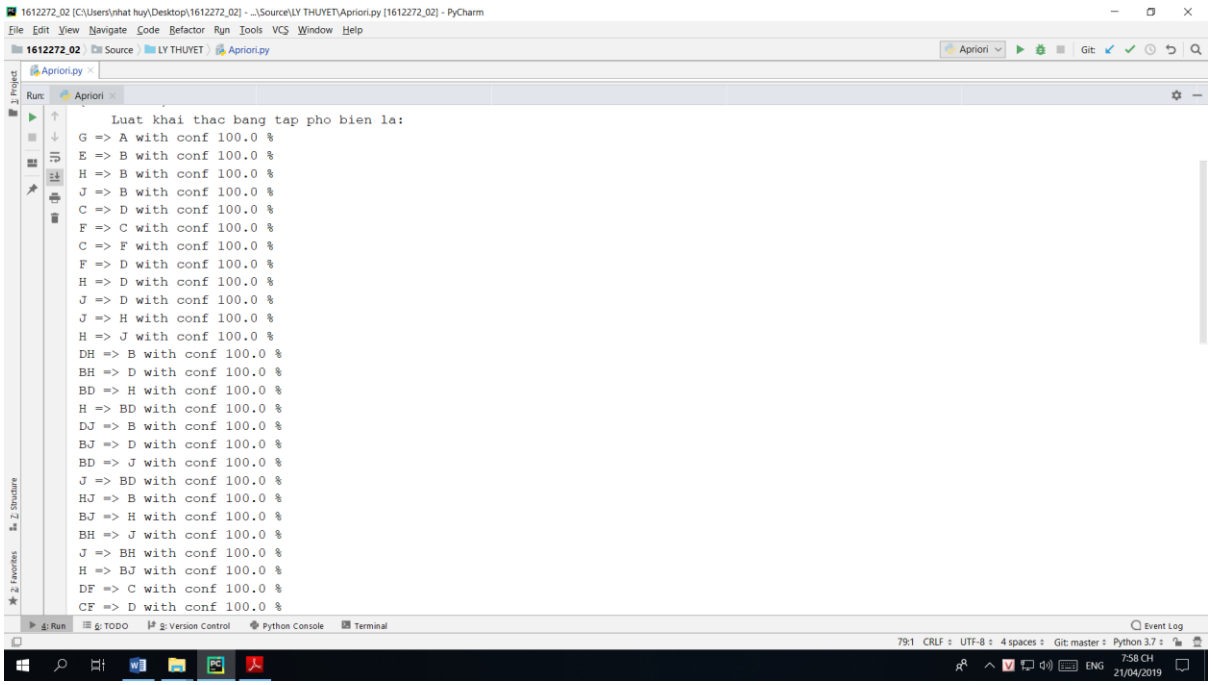


Figure 4: Luật khai thác bằng tập phổ biến

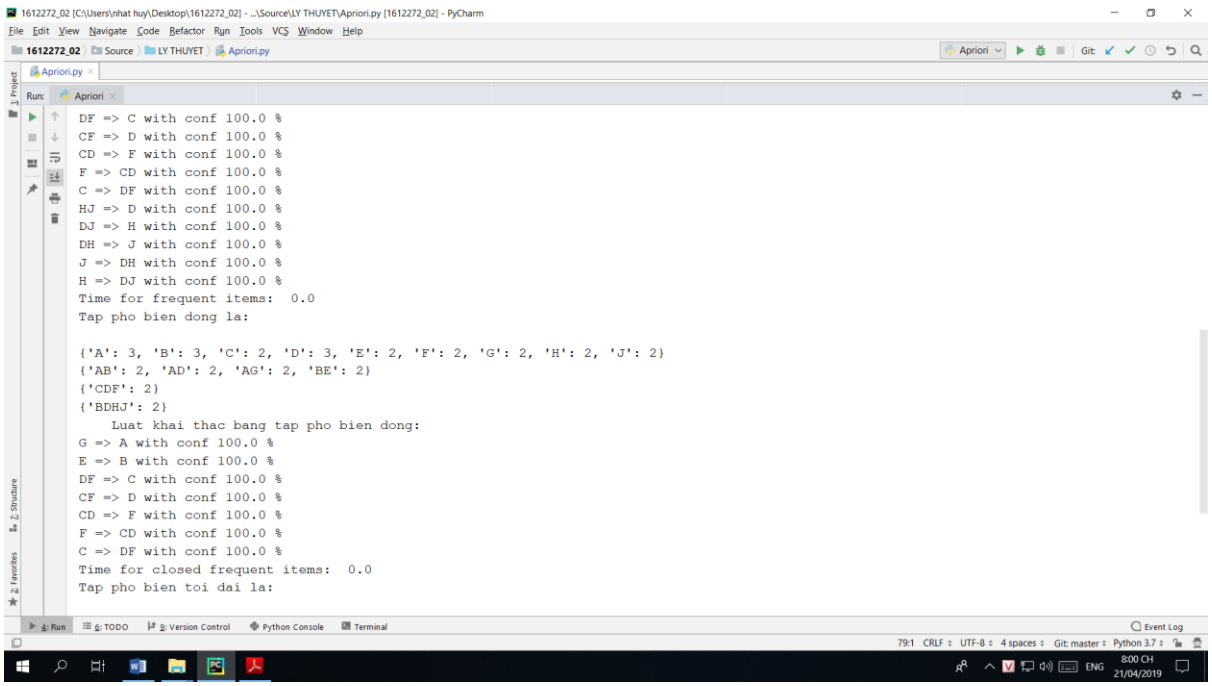


Figure 5: Luật khai thác bằng tập phổ biến (tt)

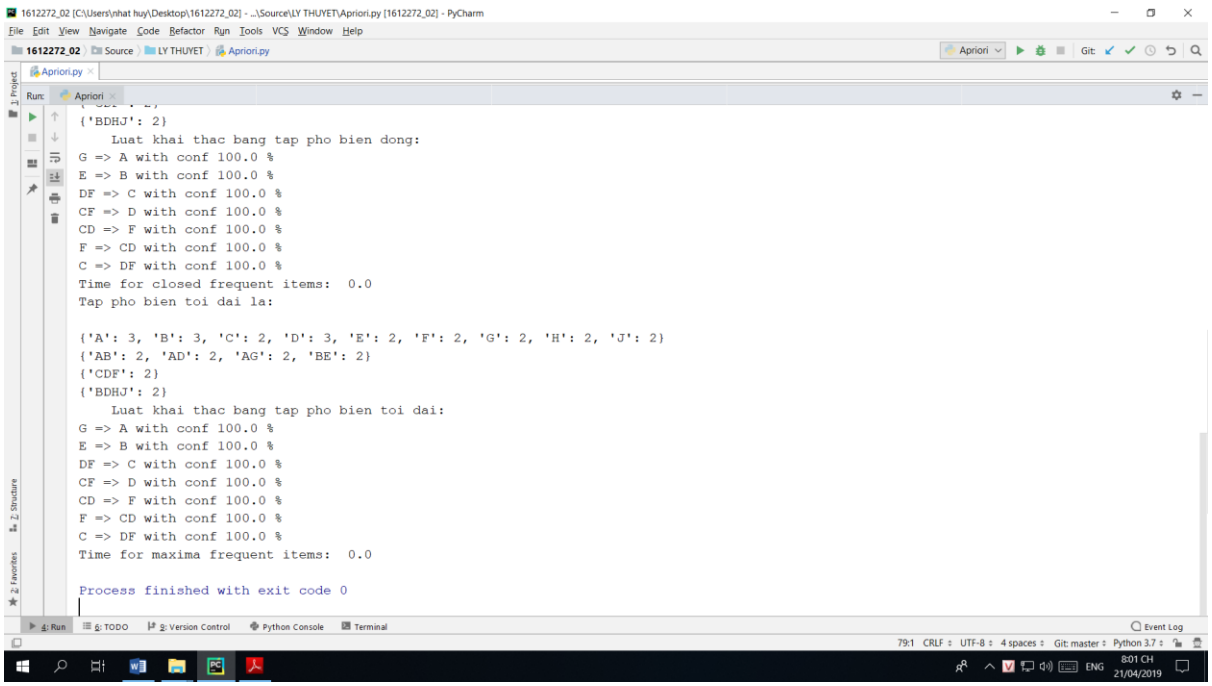


Figure 6: Luật khai thác từ tập phổ biến đóng và tối đại

Ta thấy tập luật khai thác từ tập phổ biến nhiều nhưng có những luật là con của các luật khác.

Các luật thỏa yêu cầu bài toán là:

- DF => C with conf 100.0 %
- CF => D with conf 100.0 %
- CD => F with conf 100.0 %

c. *Ứng dụng ở câu a. vào khai thác luật kết hợp. So sánh về hiệu quả*

Việc thực hiện ứng dụng khai thác luật kết hợp với tập phổ biến đóng và tập phổ biến tối đại đã thực hiện ở câu trên. Vì tập phổ biến đóng và tập phổ biến tối đại là như nhau nên luật sinh ra là như nhau. Các luật sinh ra ở tập phổ biến đóng sẽ gọn hơn tập luật sinh ra ở tập phổ biến. Về thời gian thực hiện thì ta chứ dòng "Time for …" là như nhau đối với dataset này vì tập phổ biến ít nên chúng ta không thấy sự khác biệt. Tuy nhiên, đối với dataset lớn thì thời gian giữa các phương pháp sẽ có cách biệt đáng kể

B. THỰC HÀNH

1. CÂU 1

Ta tải các file trong link được cung cấp gồm có: **plants.data, plants.names, stateabbr.txt**.

Để chuyển đổi dữ liệu ta cần xem nội dung các file

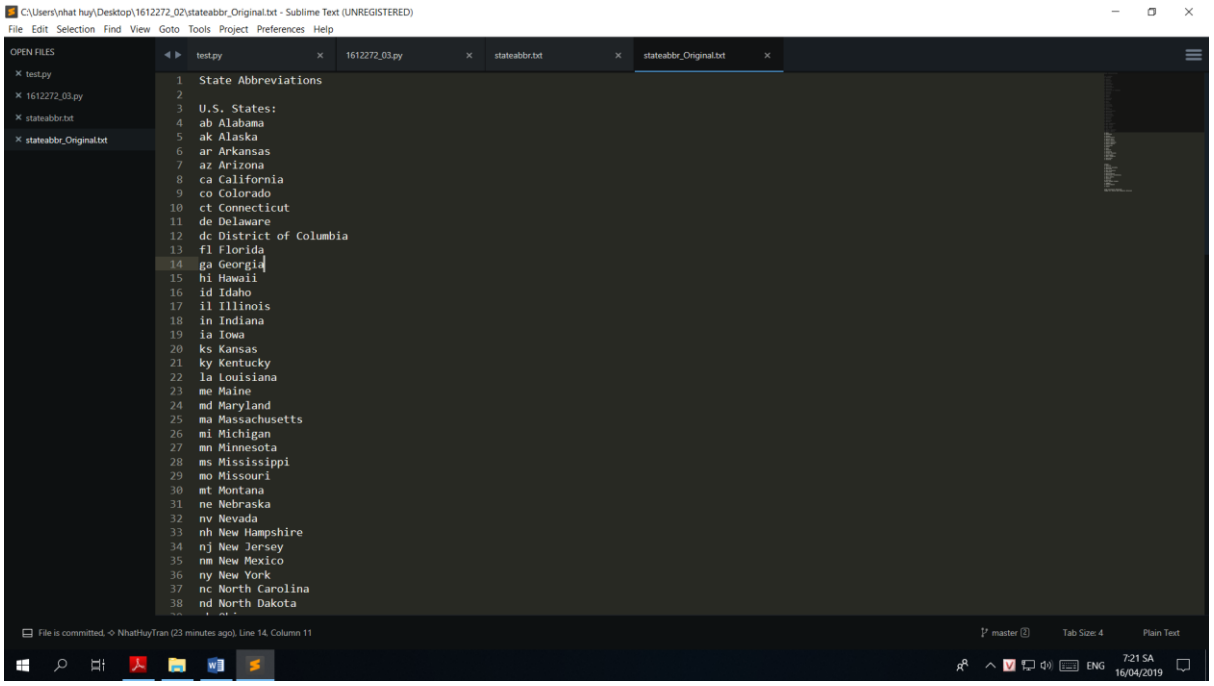


Figure 7: Nội dung file stateabbr.txt

Ta tiến hành loại bỏ các dòng không cần thiết để phục vụ cho các xử lý tiếp theo. File sau khi xử lý sẽ được lưu thành **stateabbr_process.txt**. Trong file **stateabbr.txt** có **state Prince Edward Island** không có ký hiệu viết tắt tuy nhiên trong file **plants.data** lại sử dụng **pe** thay cho **Prince Edward Island**

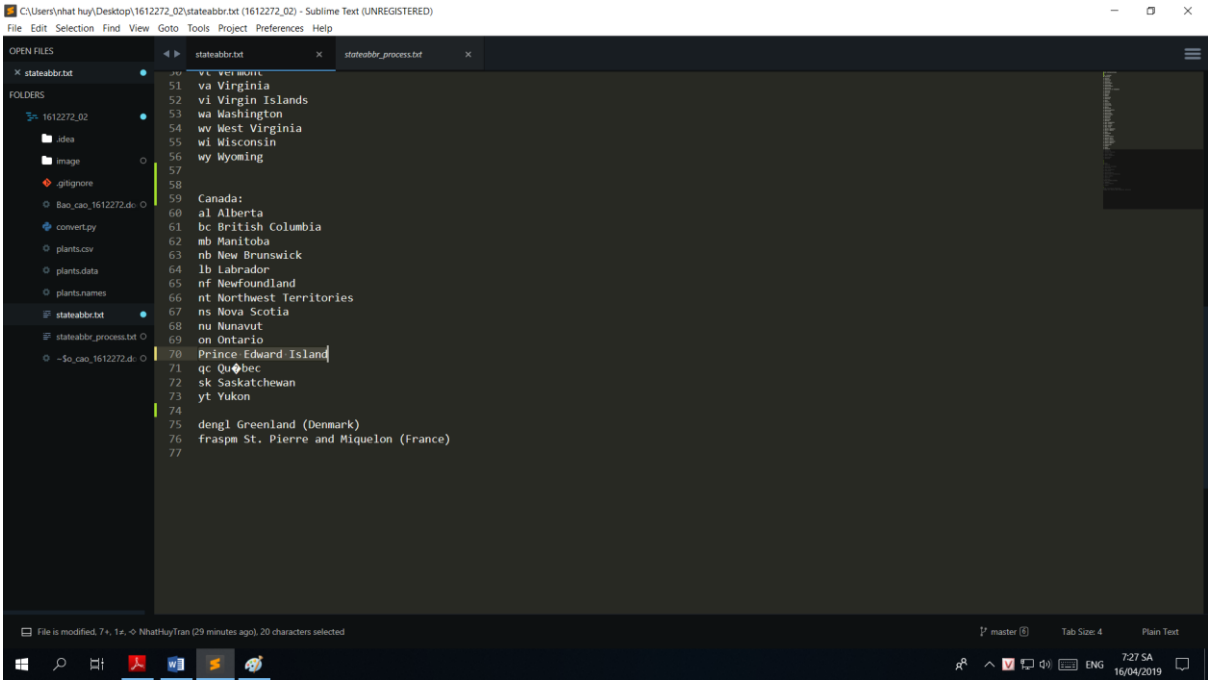


Figure 8: Thành phố Prince Edward Island trong stateabbr.txt

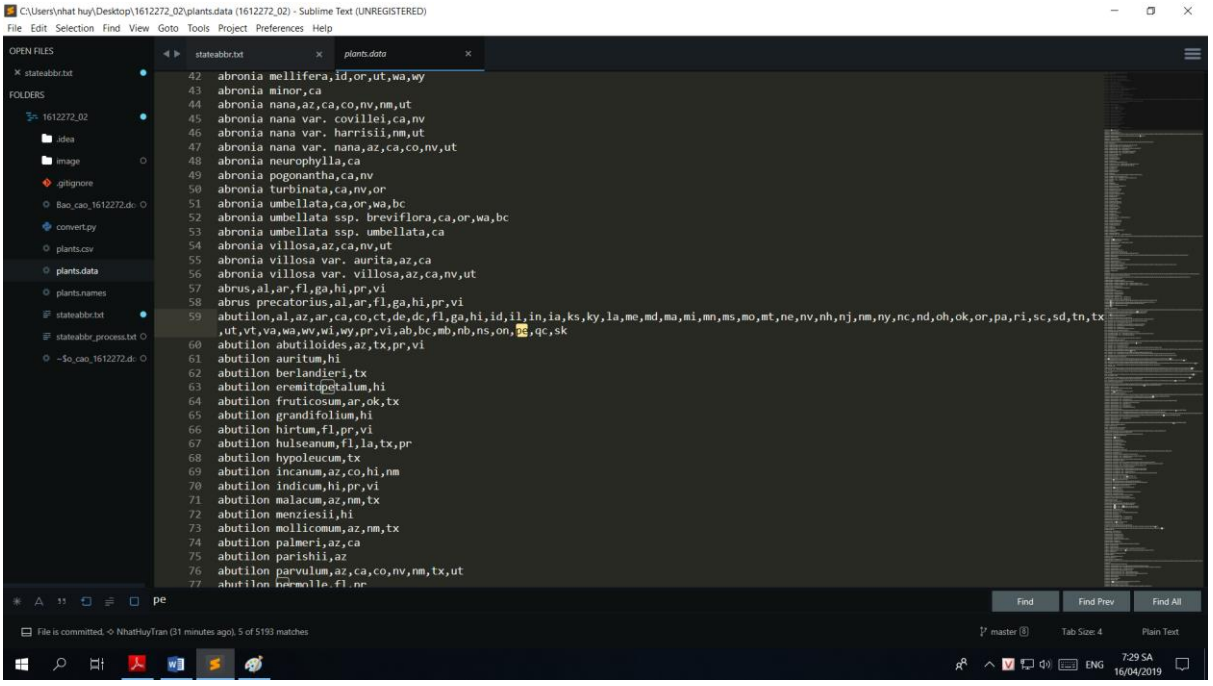


Figure 9: Ký hiệu pe được sử dụng trong plants.data

Ta sẽ thêm ký hiệu **pe** cho **Prince Edward Island** và xóa hết các dòng không cần thiết. Lưu lại thành file **stateabbr_process.txt**

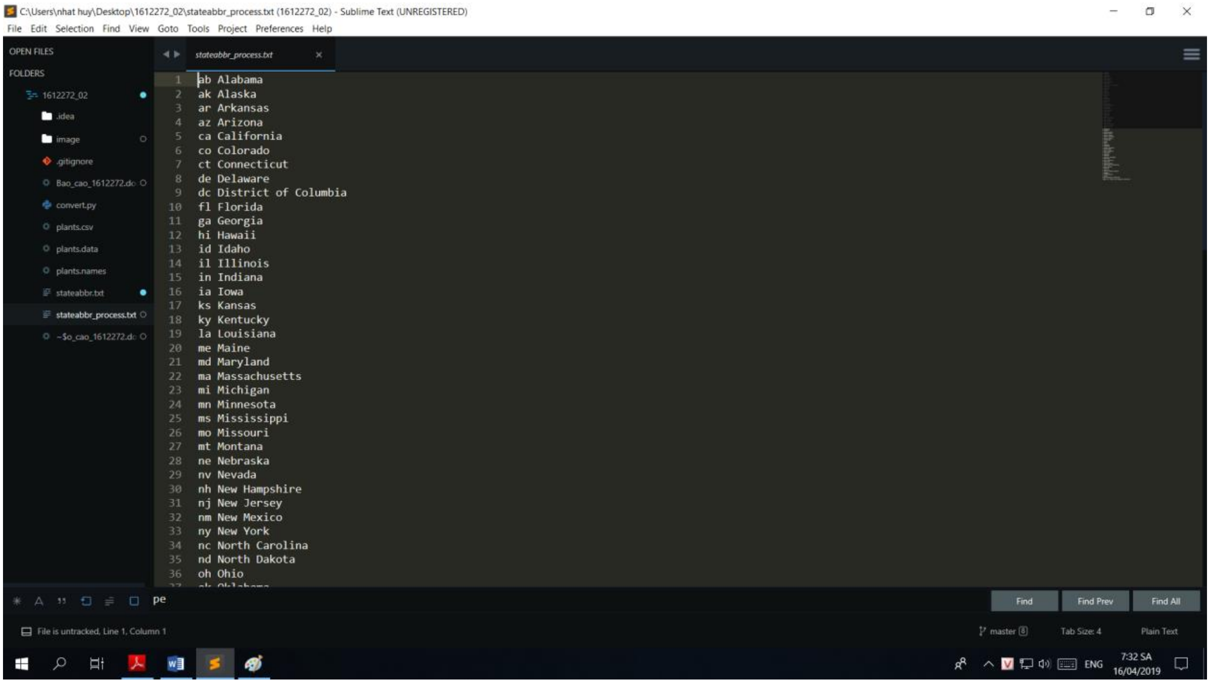


Figure 10: File stateabbr_process.txt

Để chuyển file **plants.data** thành file **plant.csv** ta chạy code file **convert.py**

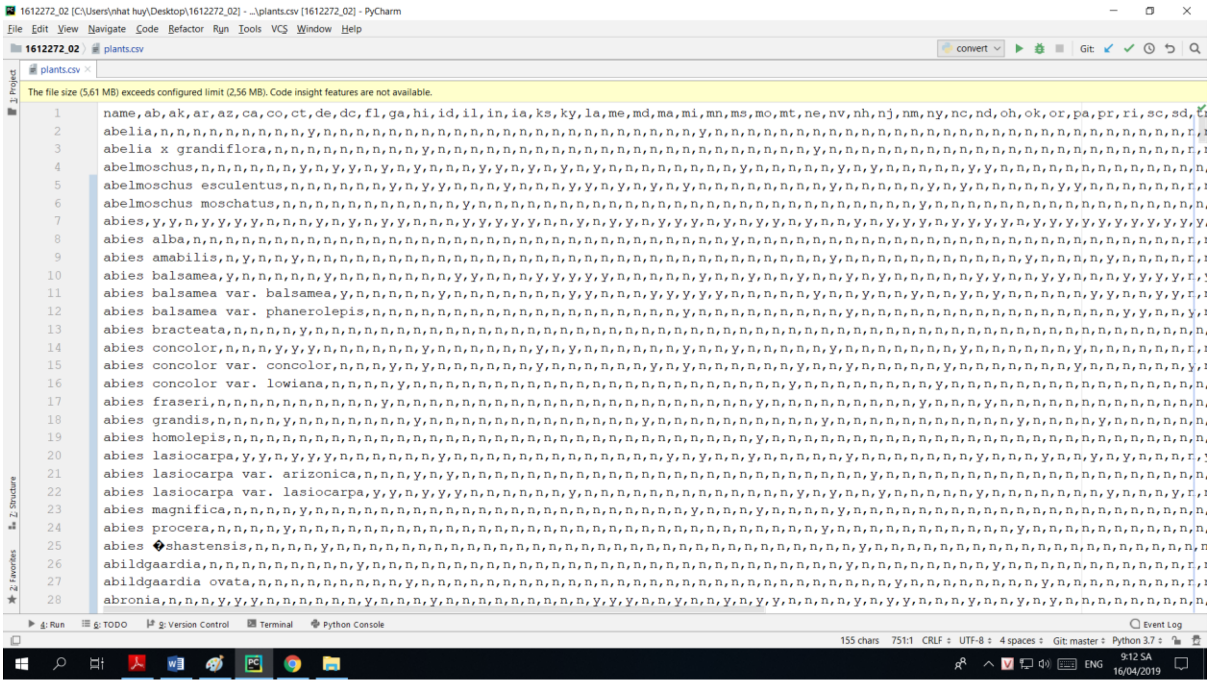


Figure 11: File plants.data sau khi chuyển

2. CÂU 2

Ta tiến hành mở file **plants.csv** bằng **Weka**

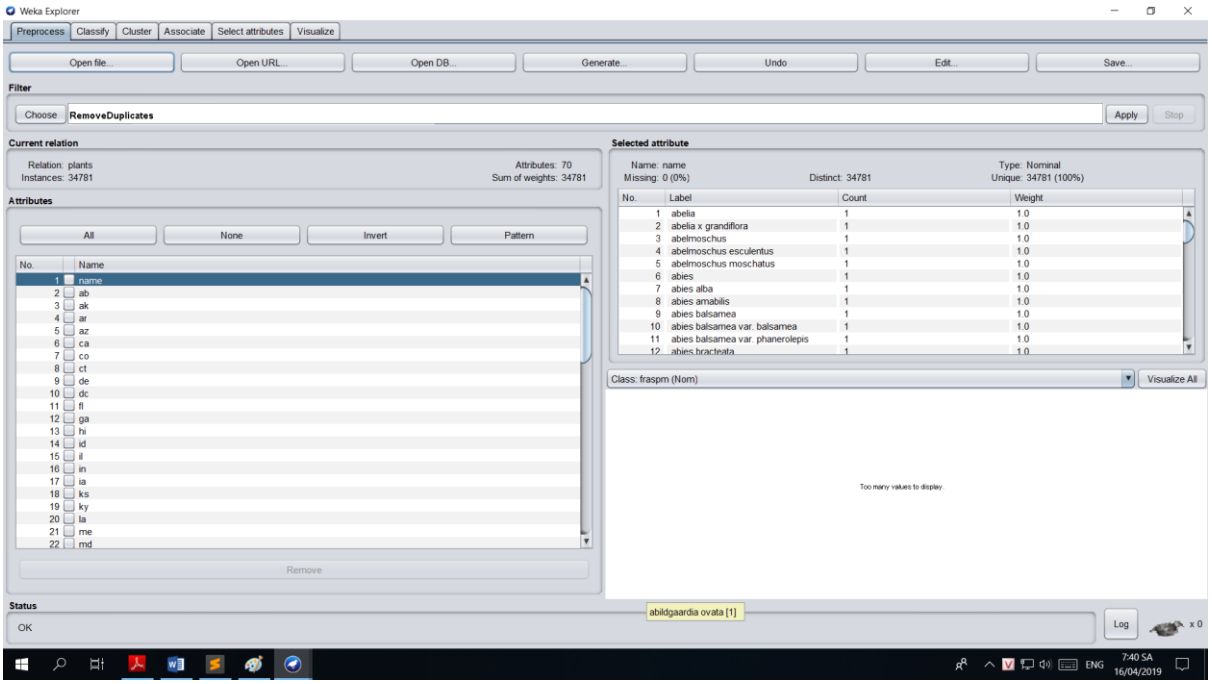


Figure 12: File plants.csv mở bằng Weka

Ta có tổng cộng 34781 cây khác nhau

Dựa vào hình ta có tổng cộng 70 thuộc tính trong đó có 1 thuộc tính name và 69 thuộc tính vùng phân bố. Vậy chúng ta có tổng cộng 69 vùng phân bố

Để xác định mỗi vùng có bao nhiêu loại cây, ta ấn vào nút **Visualize all** để xem

Figure 13: 31 vùng đầu tiên

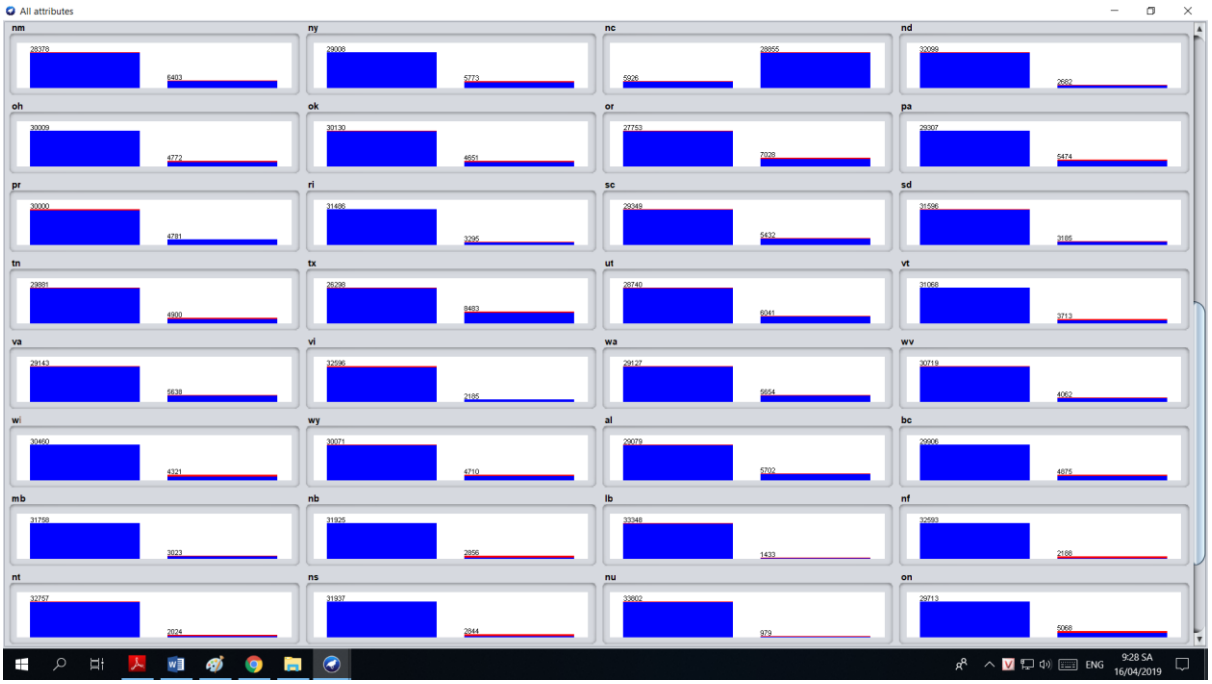


Figure 14: 32 vùng tiếp theo

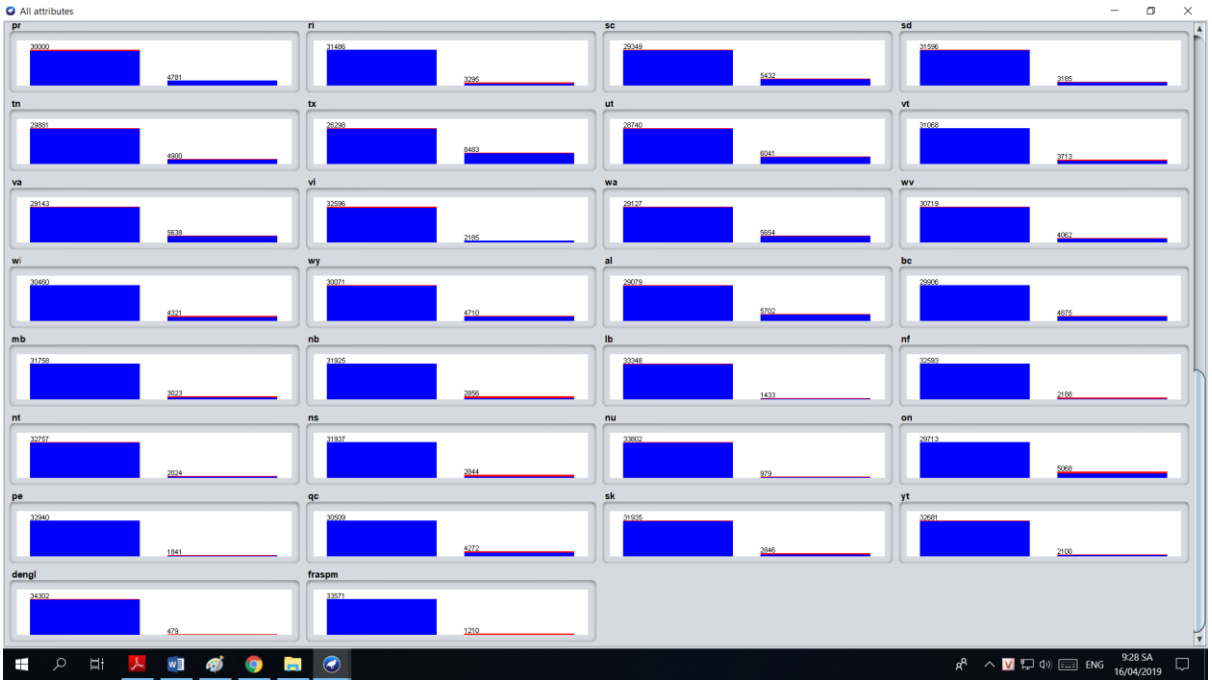
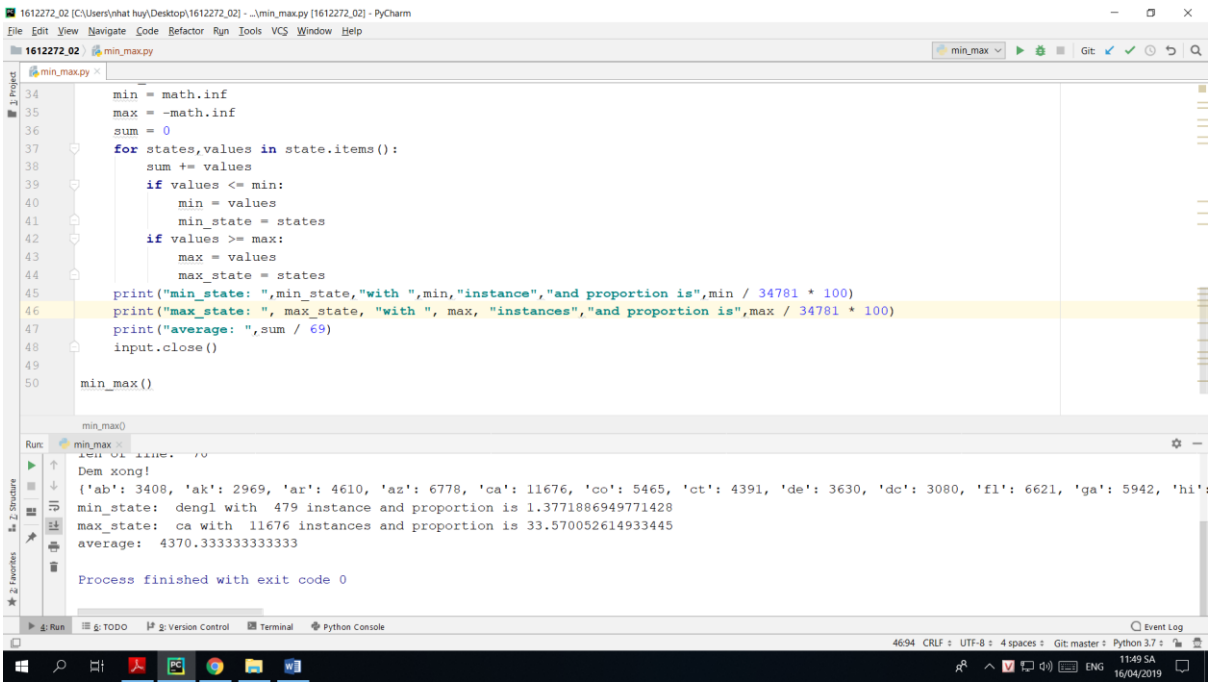


Figure 15: 6 vùng còn lại

Để kiểm tra xem vùng phân bố có ít/nhiều loài cây nhất, số lượng, tỉ lệ, trung bình một vùng phân bố bao nhiêu cây thì ta chạy file **min_max.py**



Hình 1: Kết quả khi chạy file min_max.py

Vùng ít cây nhất là **Greenland Denmark (dengl)** với 479 loại cây và tỉ lệ là 1.37%

Vùng có nhiều cây nhất là **California (ca)** với 11676 loại cây và tỉ lệ là 33.57%

Trung bình mỗi vùng có khoảng 4370 loại

3. CÂU 3

Ta thay thế các giá trị **"n"** thành **"?"** bằng cách chạy file **change_value.py**. File sau khi đã thay thế được đặt tên là **plants_changed.csv**

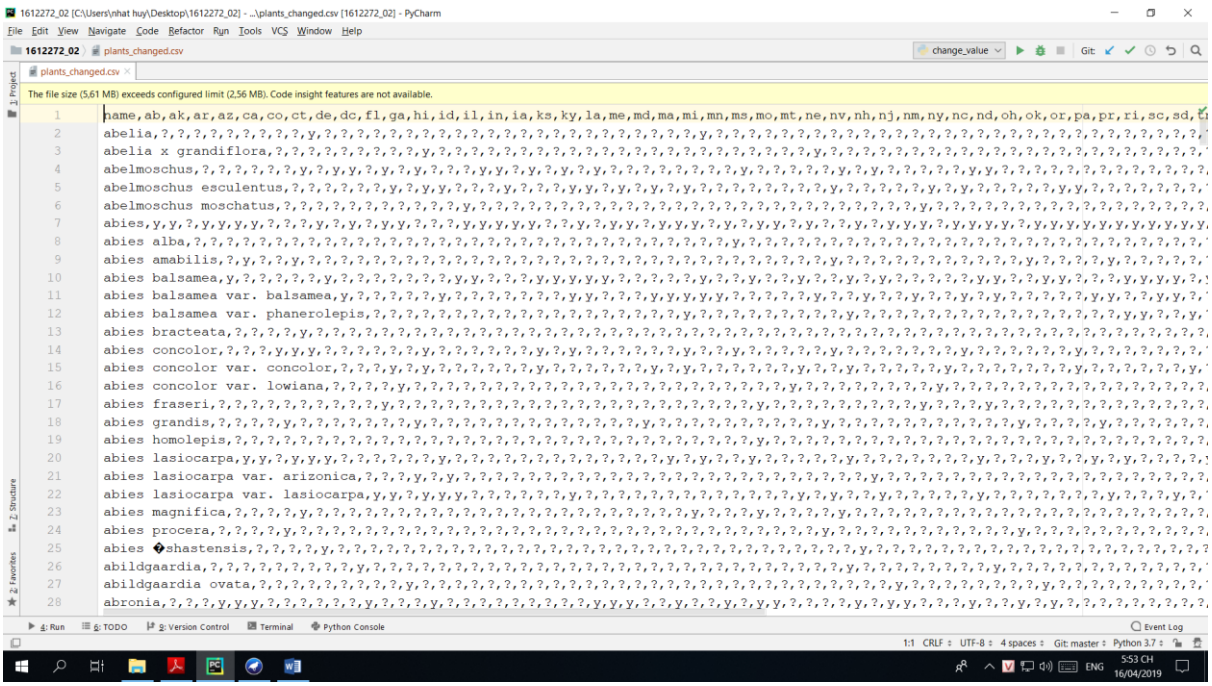


Figure 16: Nội dung file sau khi thay thế

Mở file **plants_changed.csv** bằng **Weka** và tiến hành các bước tiếp theo

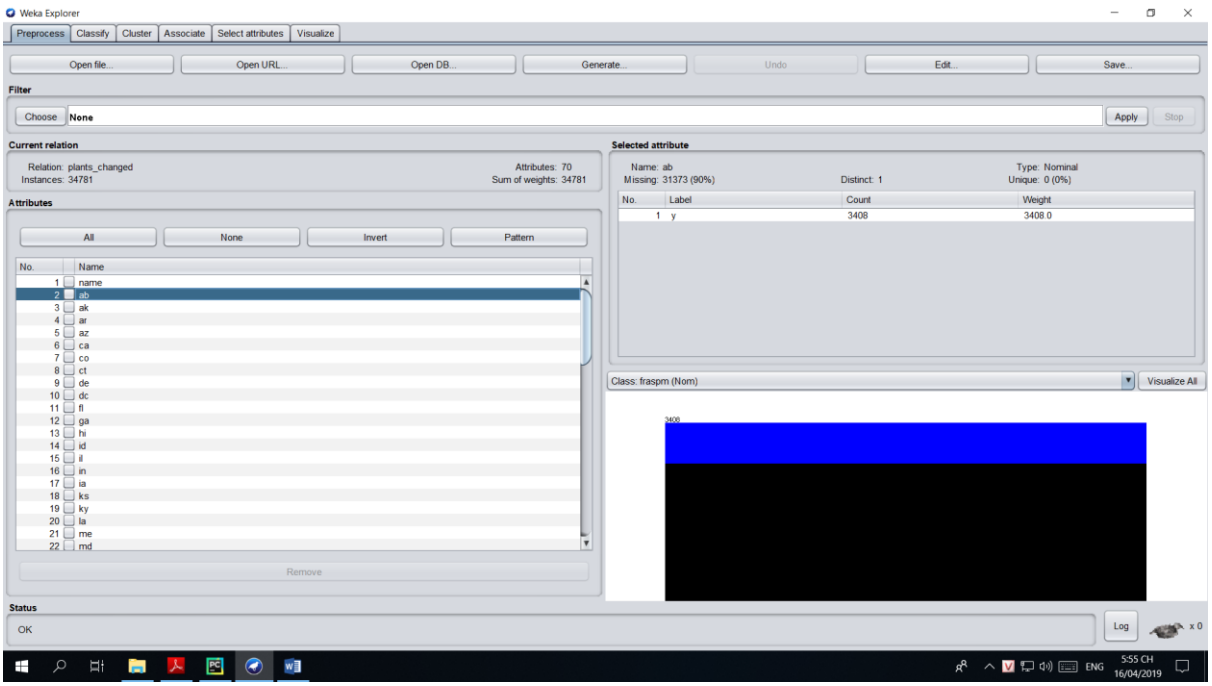


Figure 17: File plants_changed.csv được mở bằng Weka

Xóa thuộc tính name bằng cách click vào **name** và chọn **Remove**

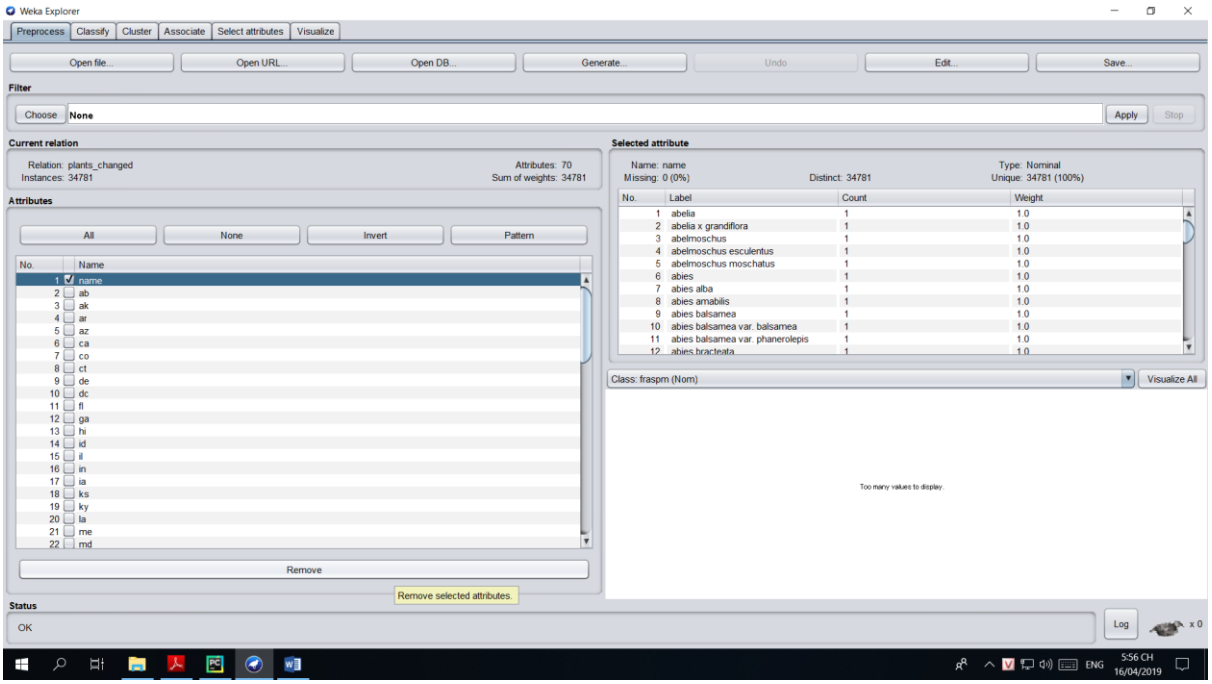


Figure 18: Chọn Remove để xóa thuộc tính name

Lưu file lại với tên là **plants.arff**

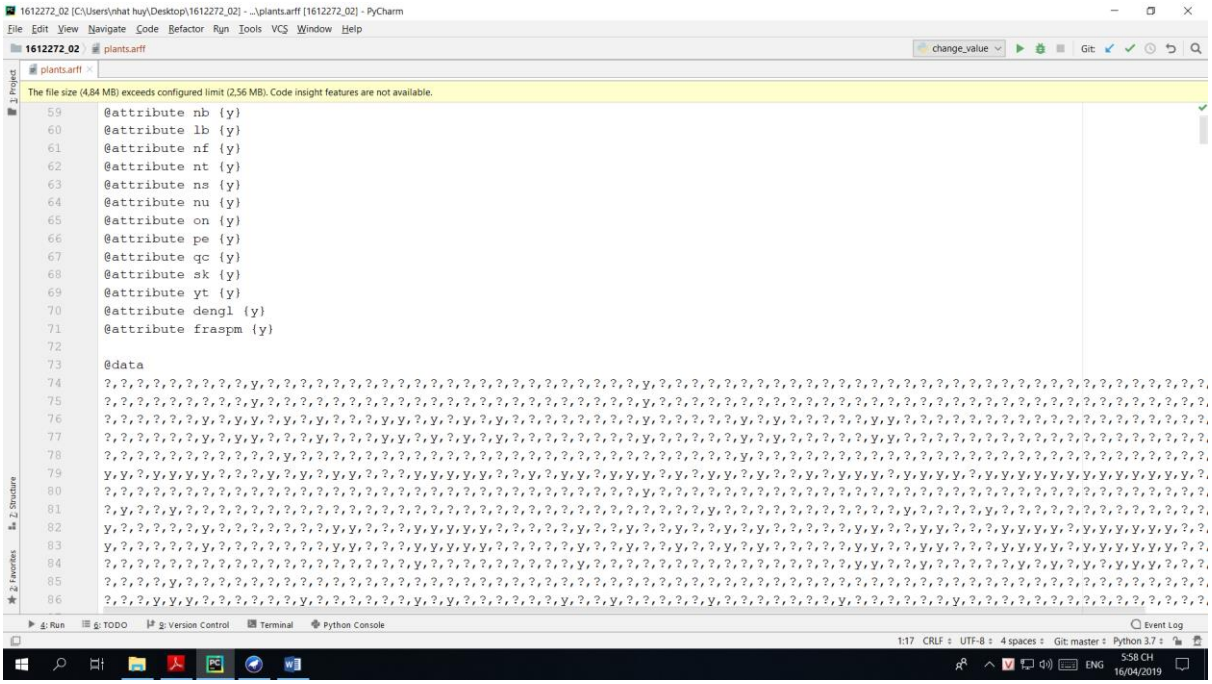


Figure 19: Nội dung file plants.arff

4. CÂU 4

Để khai thác tập phổ biến ta chuyển sang tab **Associate**

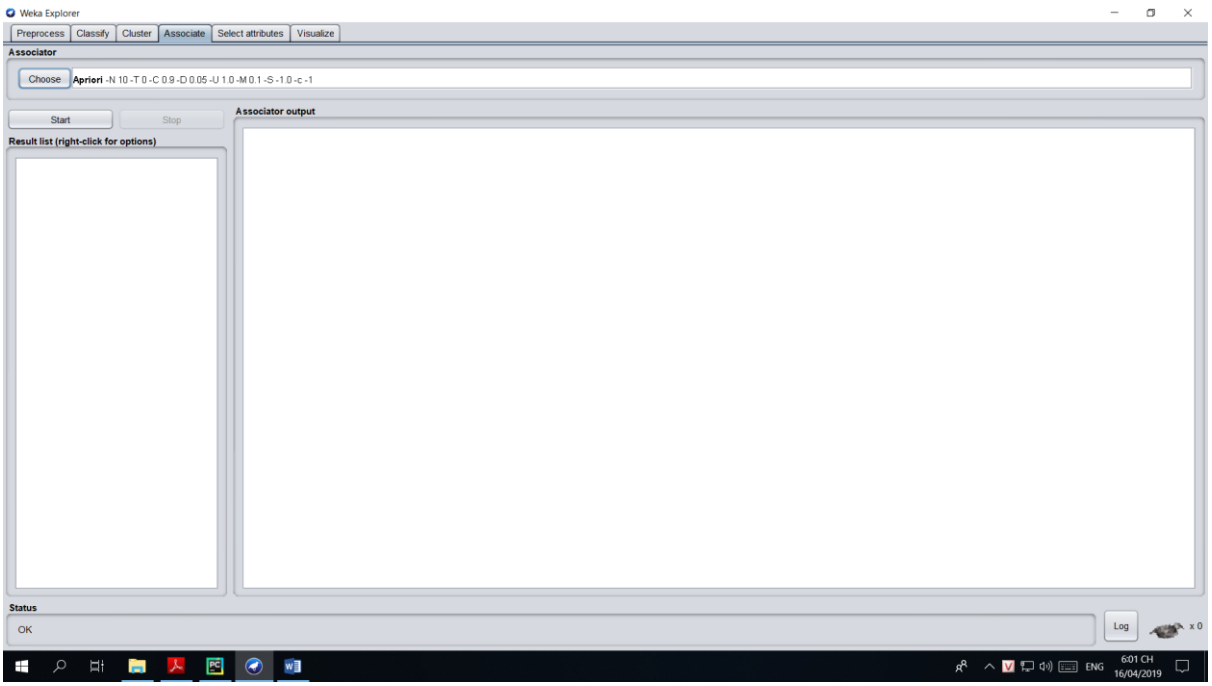


Figure 20: Tab Associate

Chọn thuật toán **Apriori** với các tham số sau

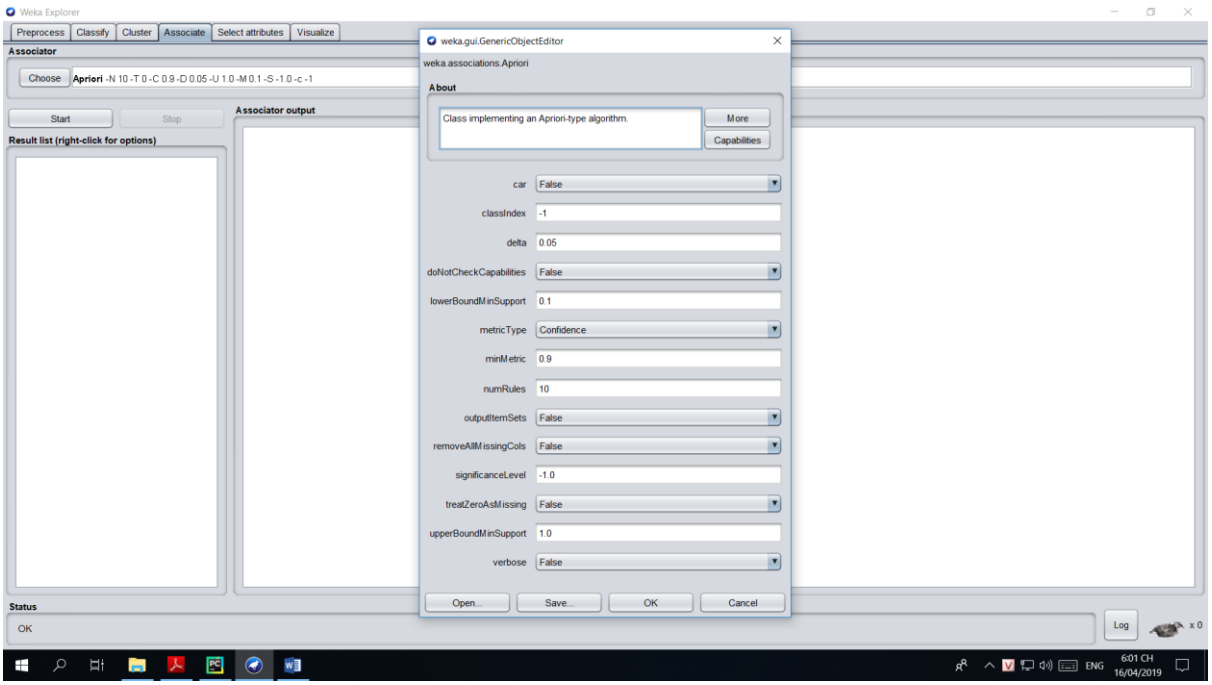


Figure 21: Thuật toán Apriori

Kết quả

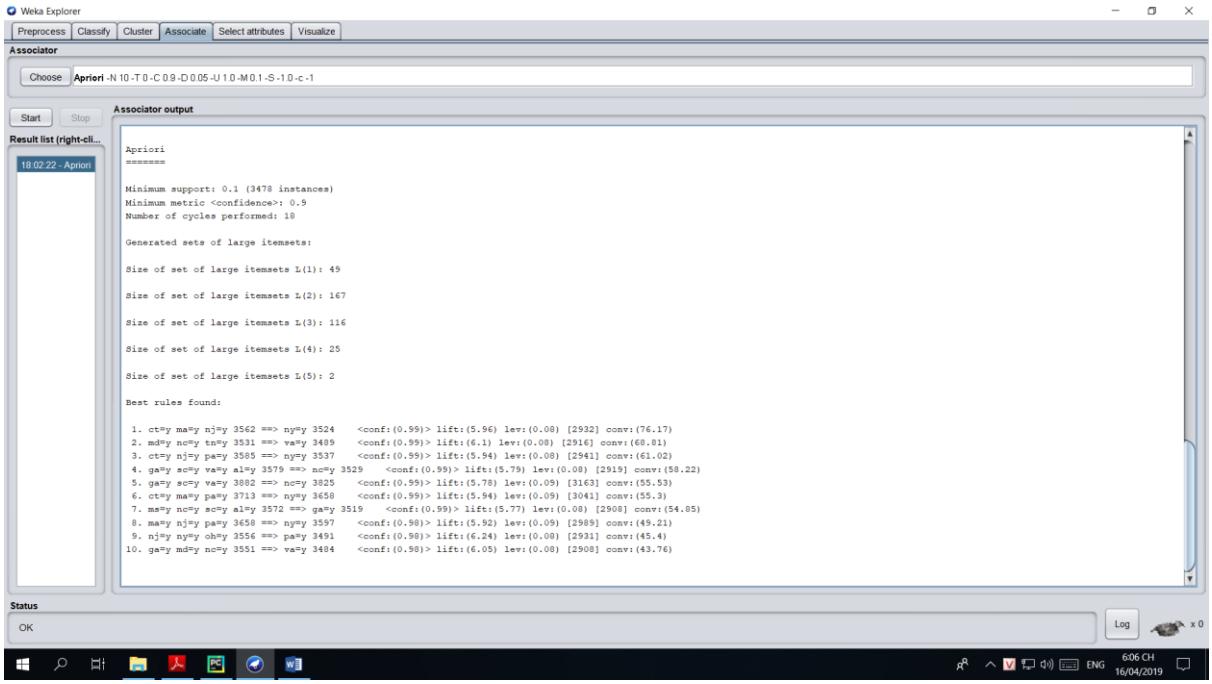


Figure 22: Kết quả chạy Apriori

KÍCH THƯỚC	SỐ LƯỢNG
1 hạng mục	49
2 hạng mục	167
3 hạng mục	116
4 hạng mục	25
5 hạng mục	2

5. CÂU 5

Chọn thuật toán **FP-Growth** với tham số **metricType** là **Confidence** và **minMetric** là **0.95**

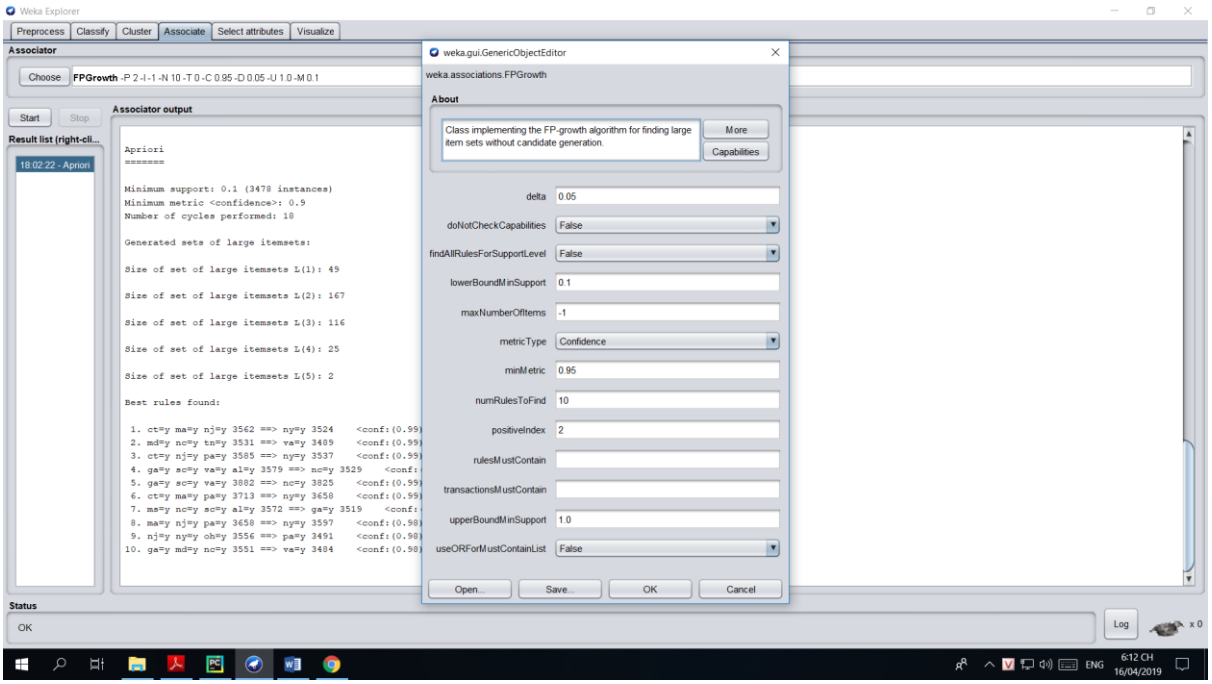


Figure 23: Tham số thuật toán FP-Growth

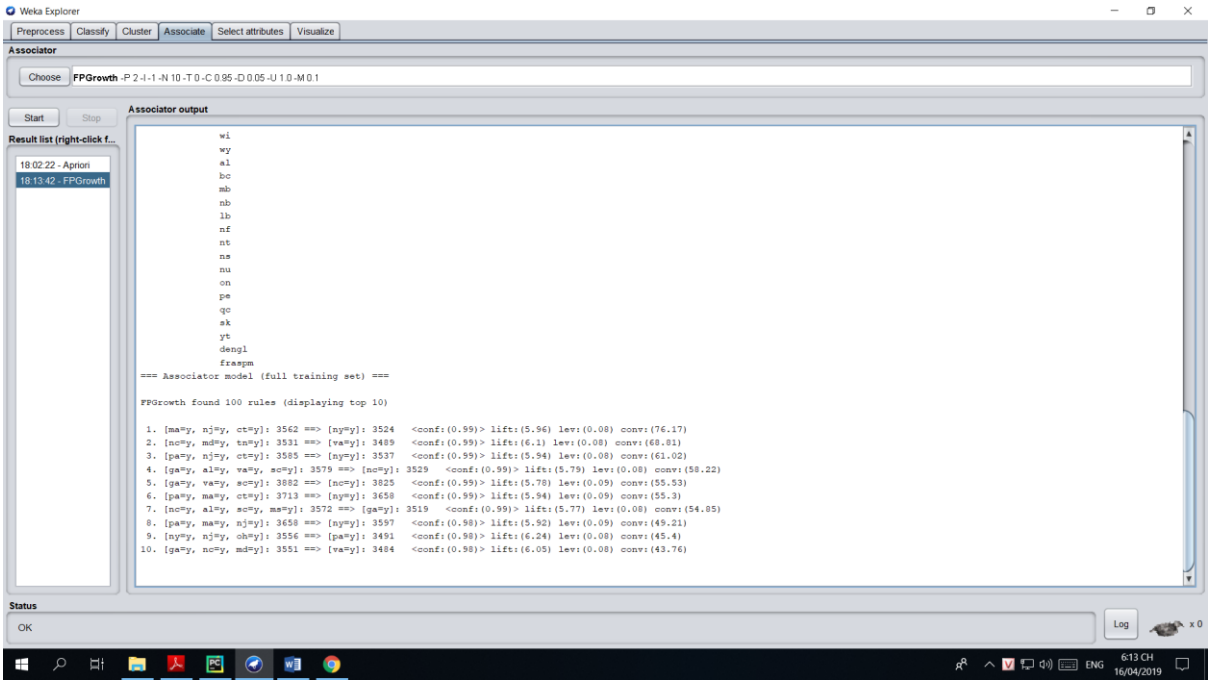


Figure 24: Kết quả chạy FP-Growth

TẬP HẠNG MỤC PHỔ BIẾN	SỐ LƯỢNG LUẬT
ma=y, nj=y, ct=y	3562
nc=y, md=y, tn=y	3531
pa=y, nj=y, ct=y	3585
ga=y, al=y, va=y, sc=y	3579
ga=y, va=y, sc=y	3882
pa=y, ma=y, ct=y	3713
nc=y, al=y, sc=y, ms=y	3572
pa=y, ma=y, nj=y	3658
ny=y, nj=y, oh=y	3556
ga=y, nc=y, md=y	3551