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

**🙥🟊🙧**

**BÀI TẬP 2: KHAI THÁC LUẬT KẾT HỢP**

1. **THÔNG TIN SINH VIÊN**

Họ và tên: **TRẦN NHẬT HUY**

Mssv: **1612272**

Email: [**nhathuy13598@gmail.com**](mailto:nhathuy13598@gmail.com)

Sđt: **0354 878 677**

1. **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% |  |

1. **CHI TIẾT BÀI LÀM**
2. **LÝ THUYẾT**
3. 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

1. Cài đặt các thuật toán
2. 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 và FP\_Growth.py

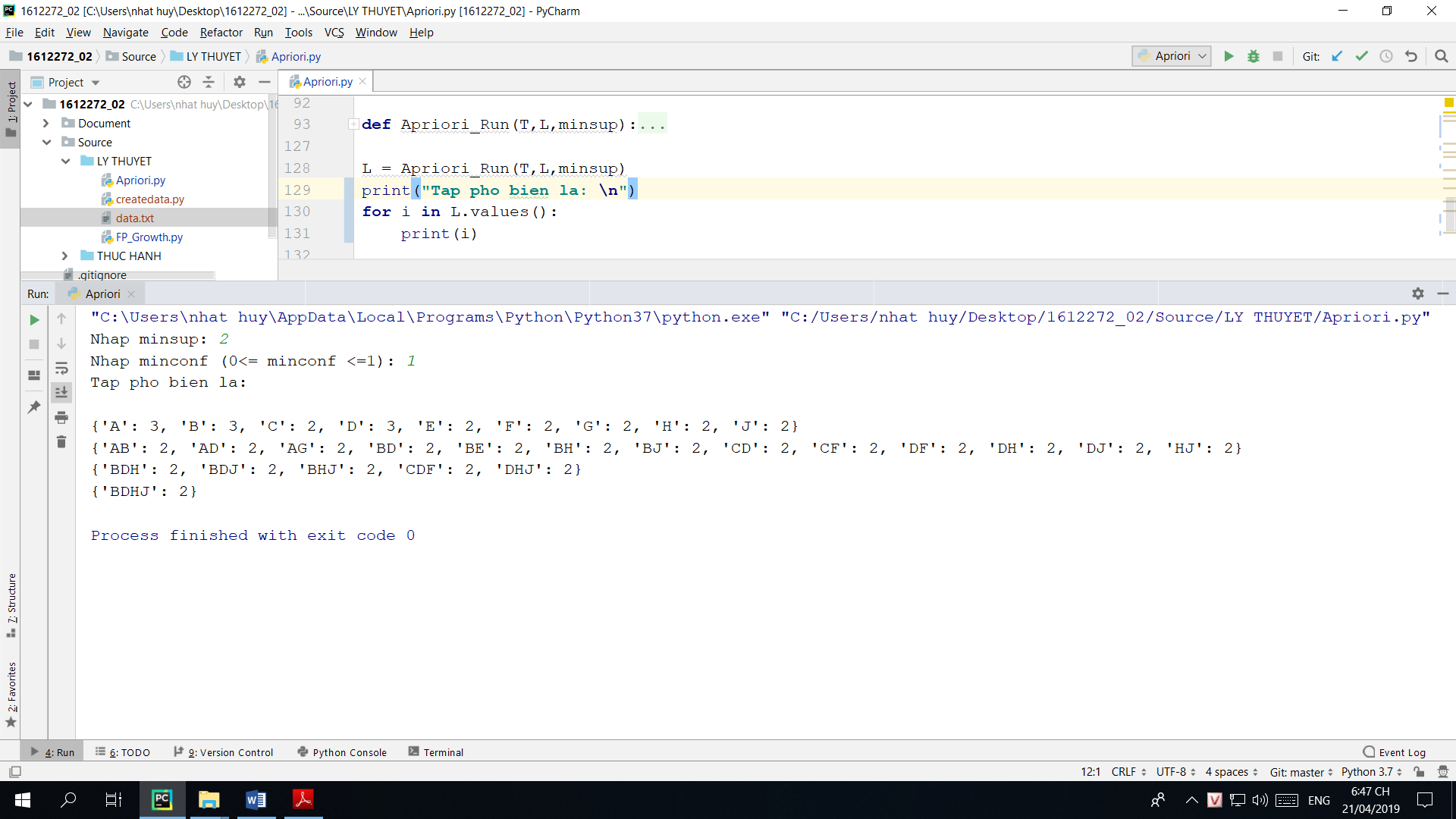


Figure 1: Tập phổ biến khi sử dụng Apriori.py

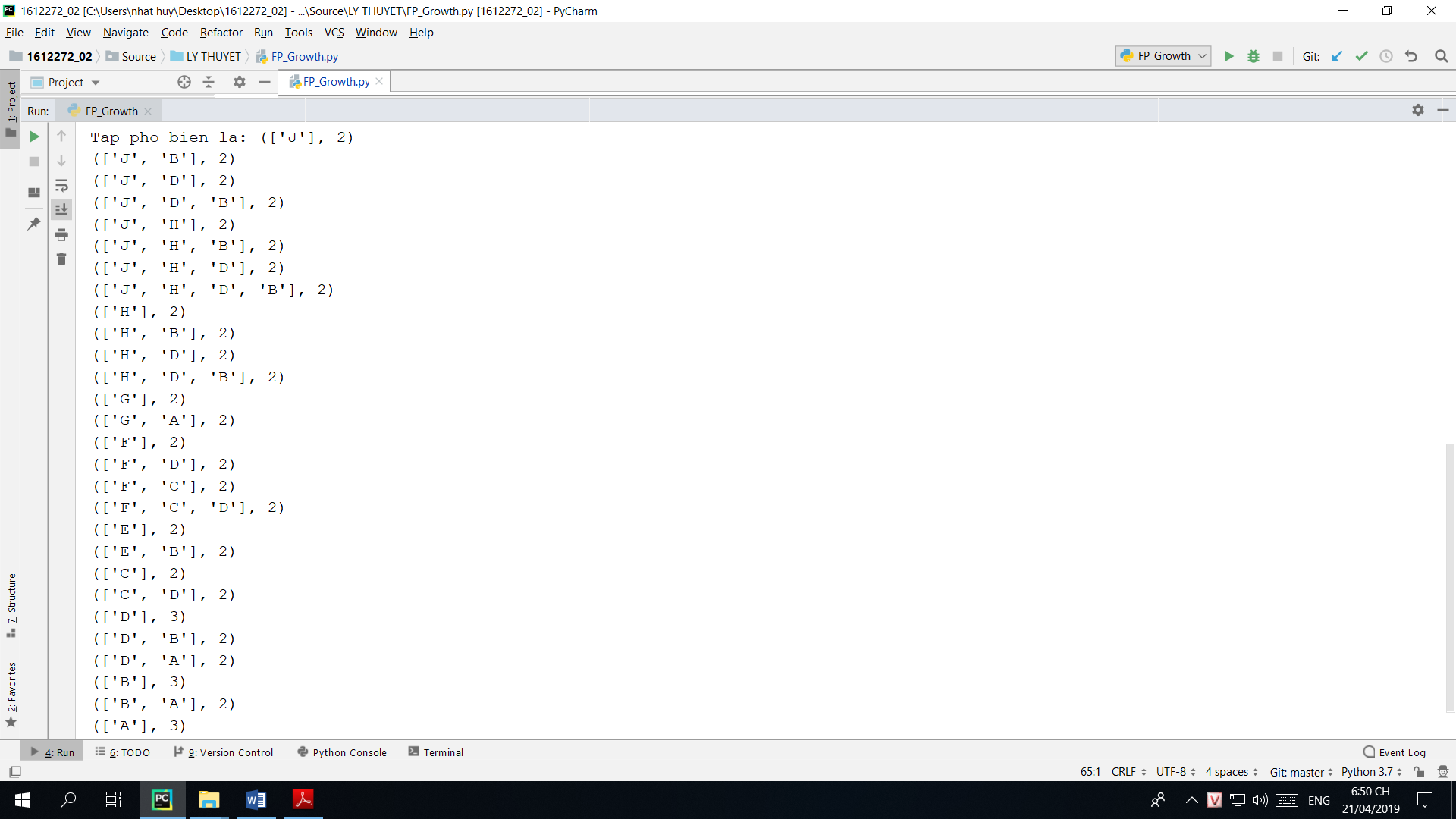


Figure 2: Tập phổ biến khi chạy FP\_Growth.py

Kết quả của 2 phương pháp này là giống nhau

|  |  |
| --- | --- |
| k-item set | Item |
| 1-item set |  |
| 2-item set |  |
| 3-item set |  |
| 4-item set |  |

1. Tìm tất cả luật kết hợp thỏa minsup và minconf
2. Ứng dụng ở câu a. vào khai thác luật kết hợp. So sánh về hiệu quả
3. **THỰC HÀNH**
4. **CÂU 1**

Ta tải các file trong [link](http://archive.ics.uci.edu/ml/machine-learning-databases/plants/) được cung cấp gồm có: plants.data, plants.names, stateabbr.txt.

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

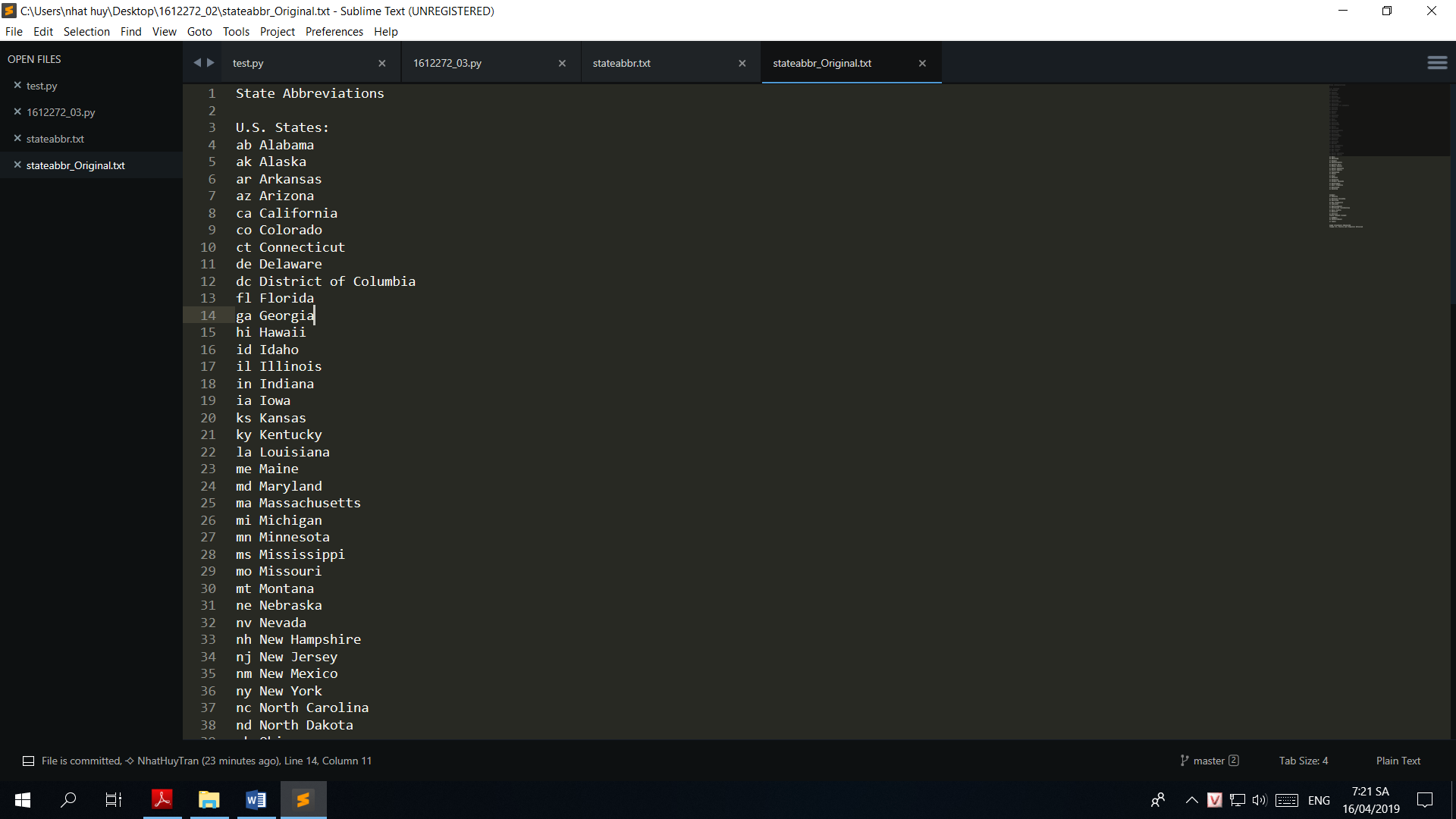


Figure 3: Nội dung file stateabbr.txt

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

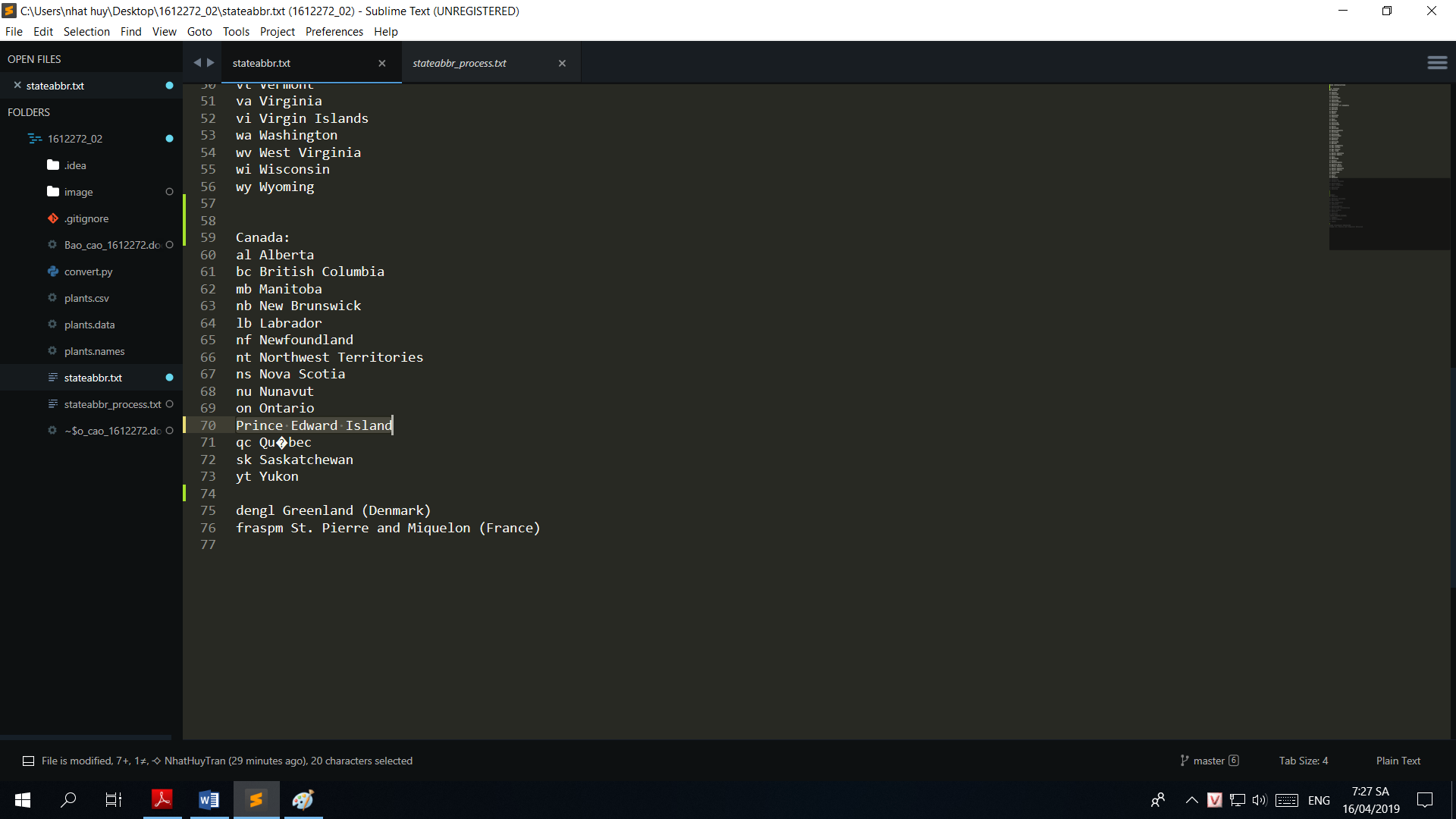


Figure 4: Thành phố Prince Edward Island trong stateabbr.txt

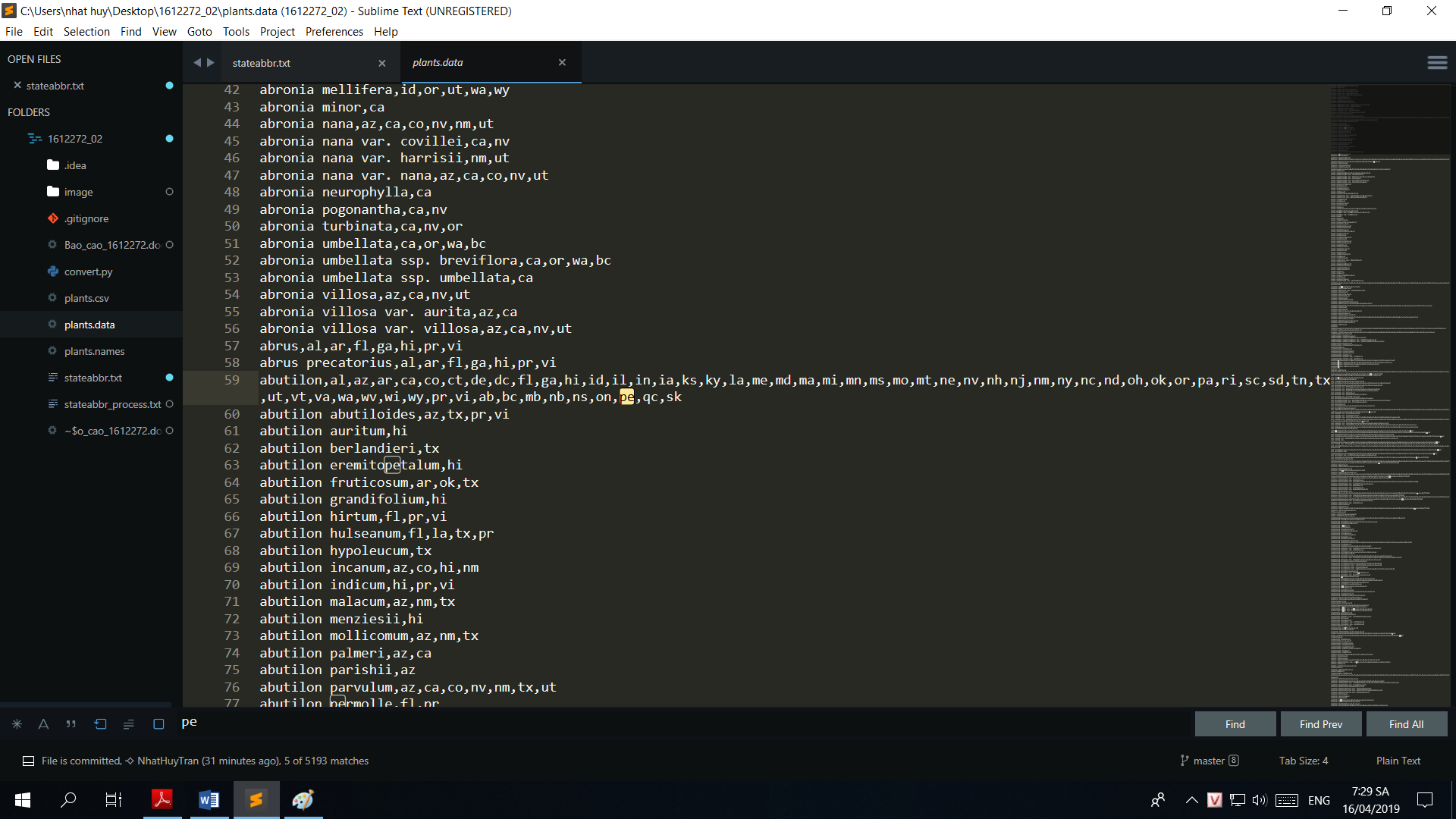


Figure 5: Ký hiệu pe được sử dụng trong plants.data

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

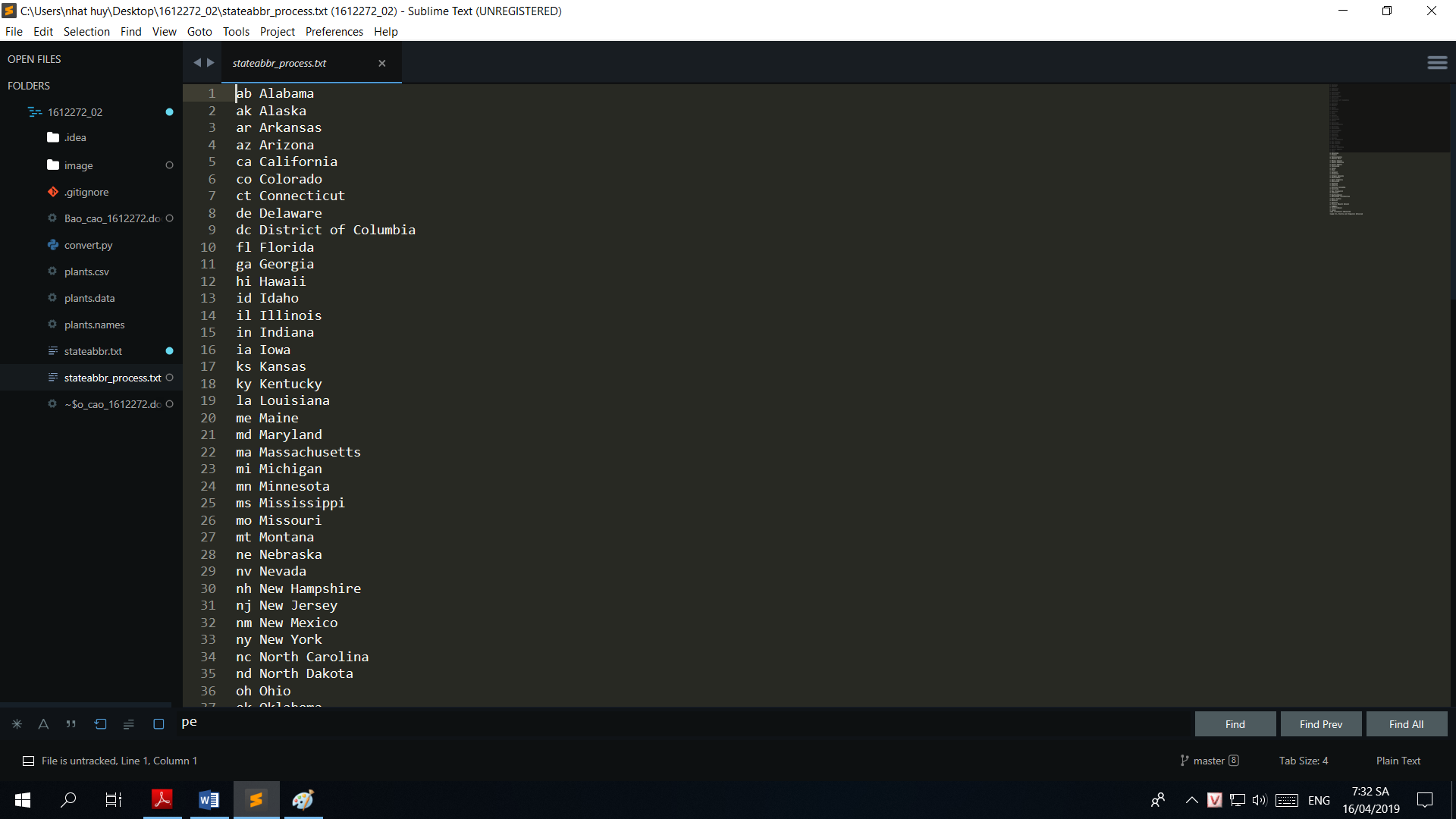


Figure 6: File stateabbr\_process.txt

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

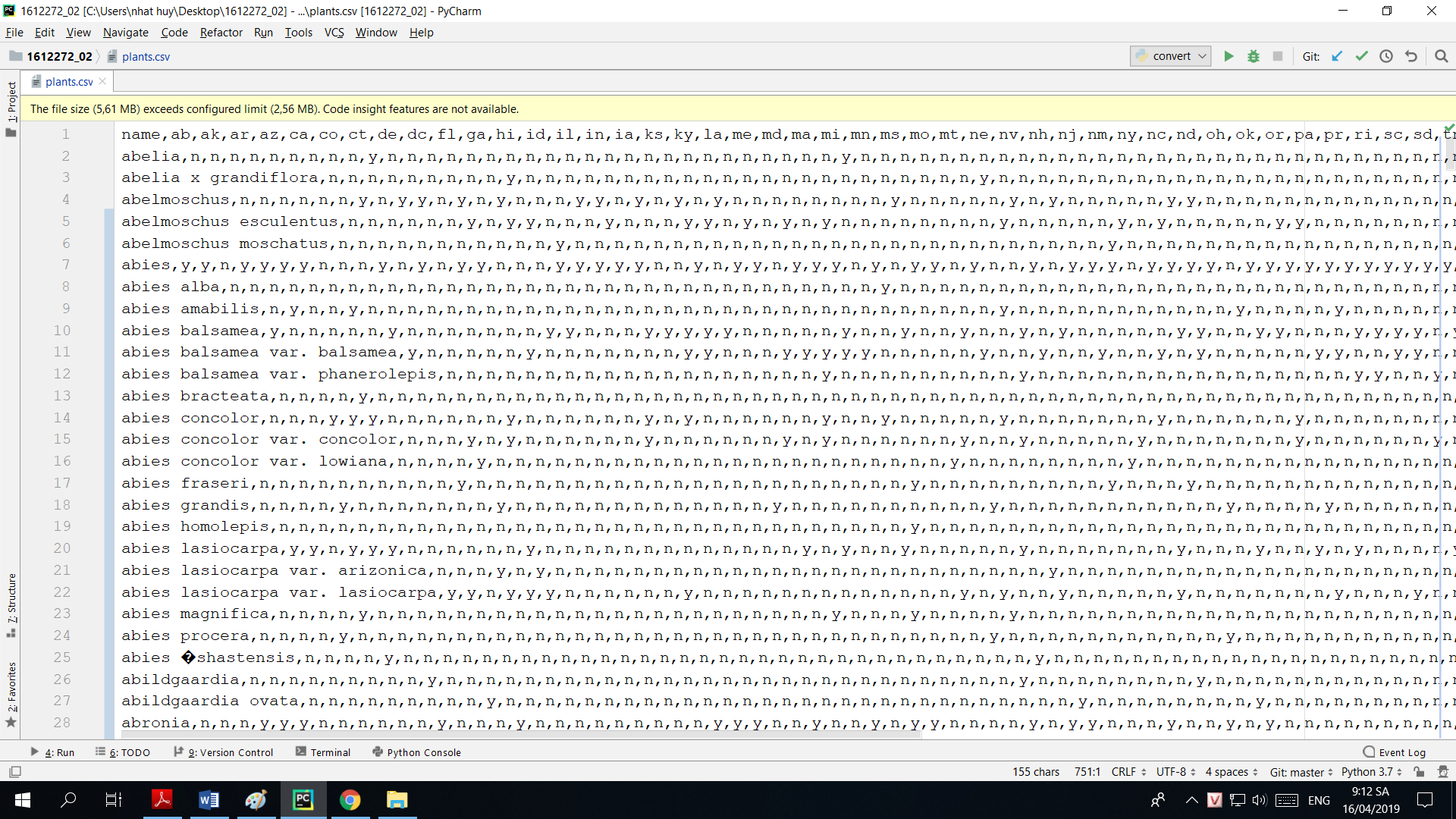


Figure 7: File plants.data sau khi chuyển

1. **CÂU 2**

Ta tiến hành mở file plants.csv bằng Weka

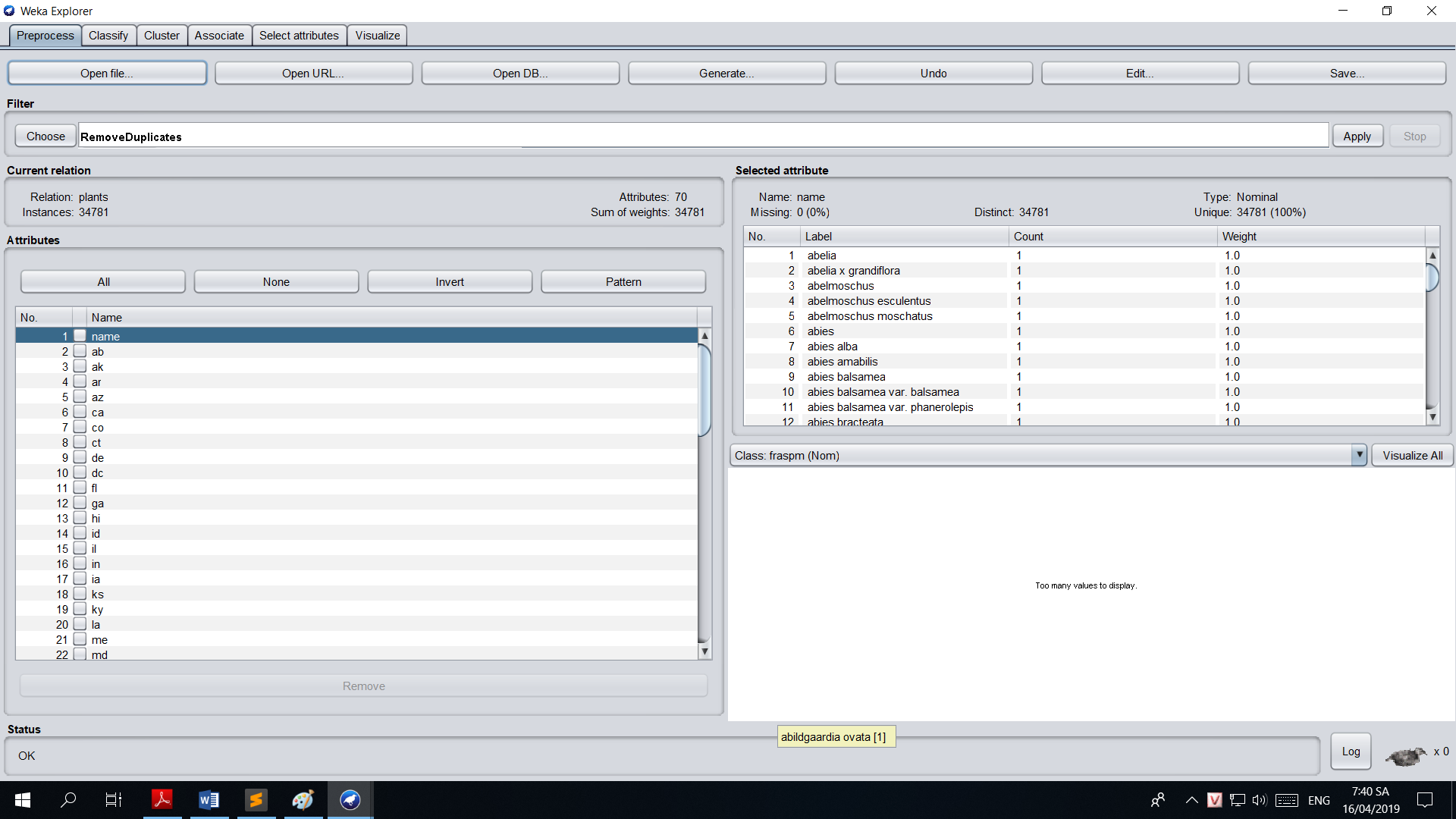


Figure 8: File plants.csv mở bằng Weka

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 9: 31 vùng đầu tiên

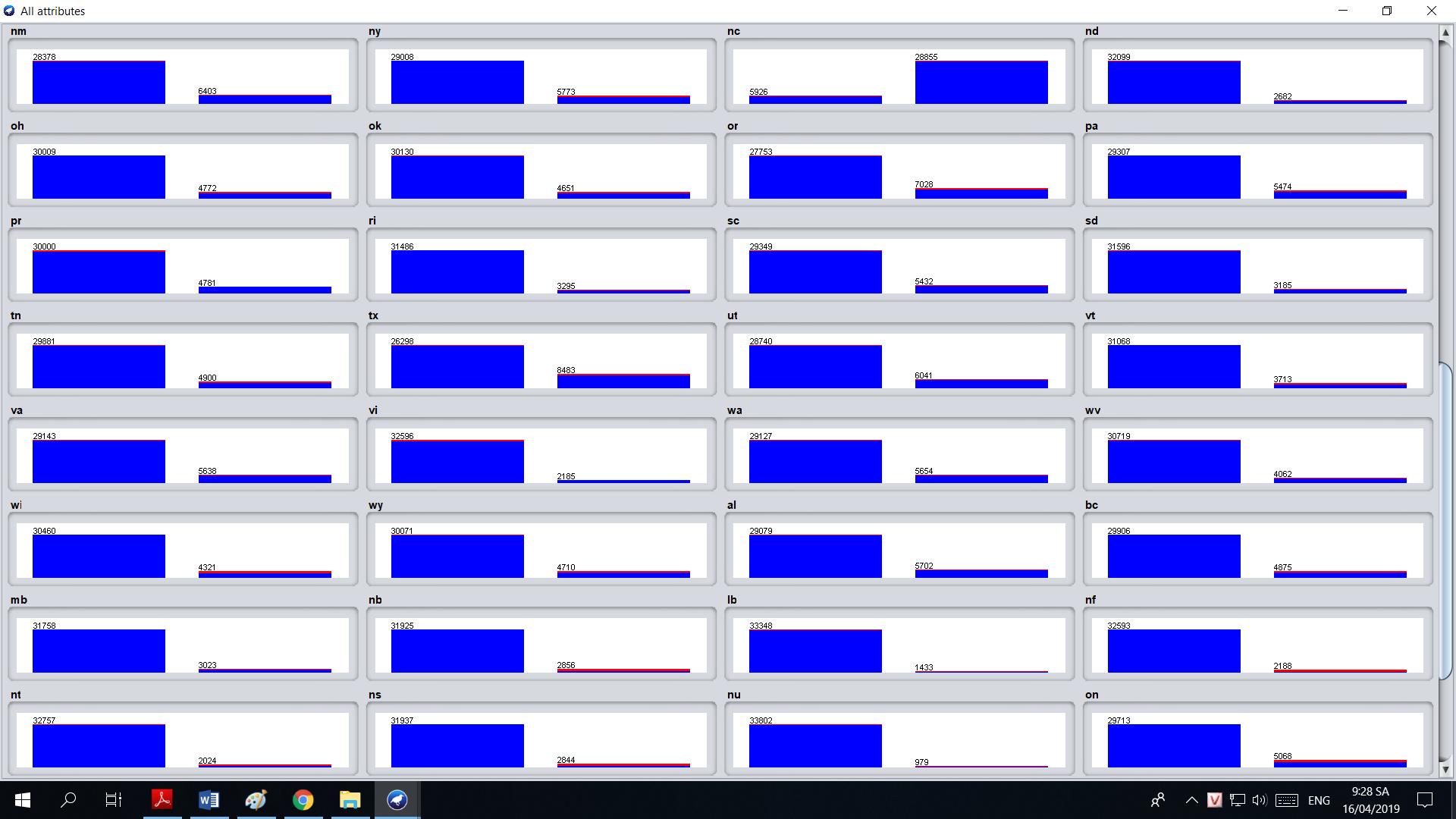


Figure 10: 32 vùng tiếp theo

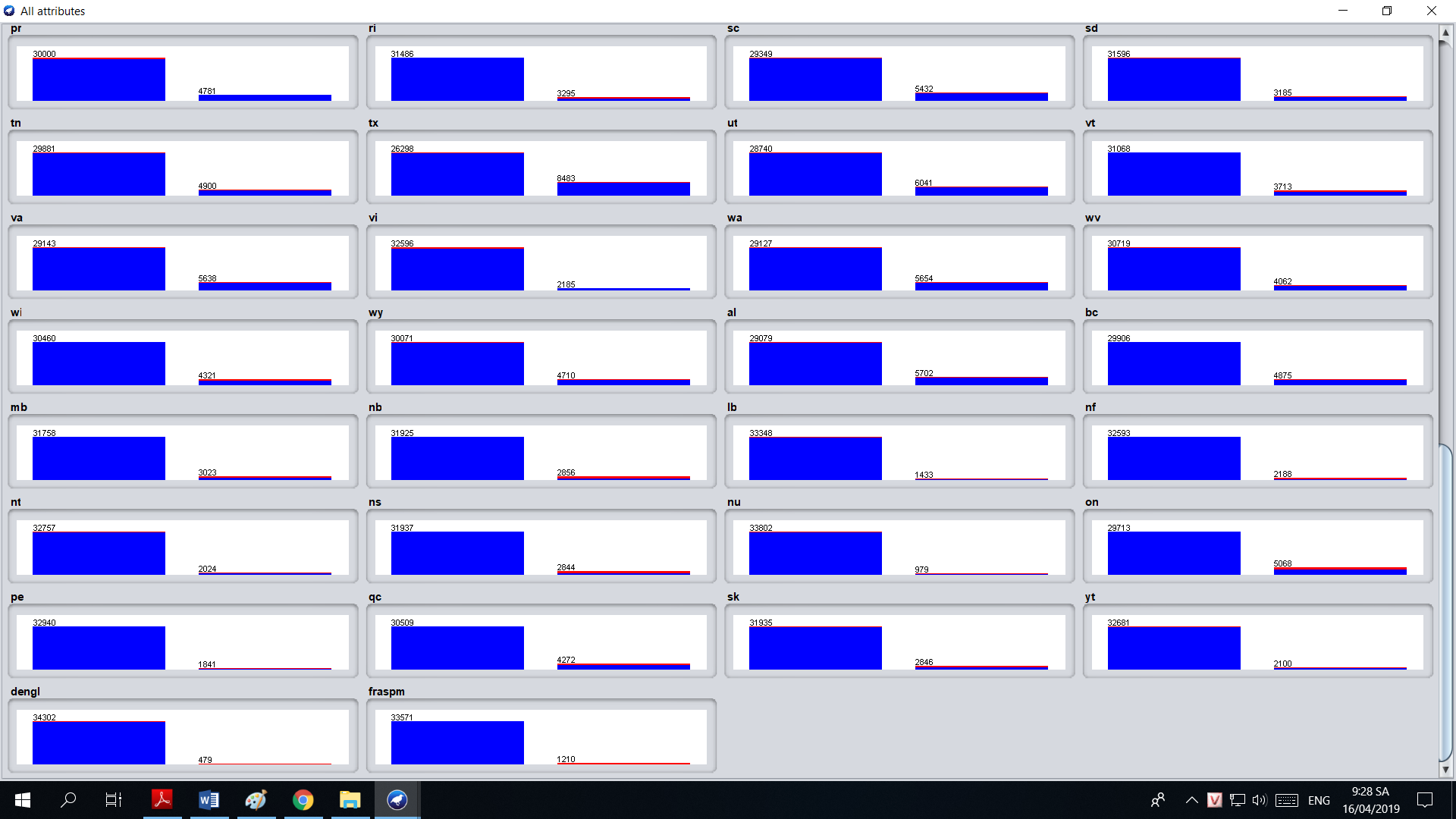
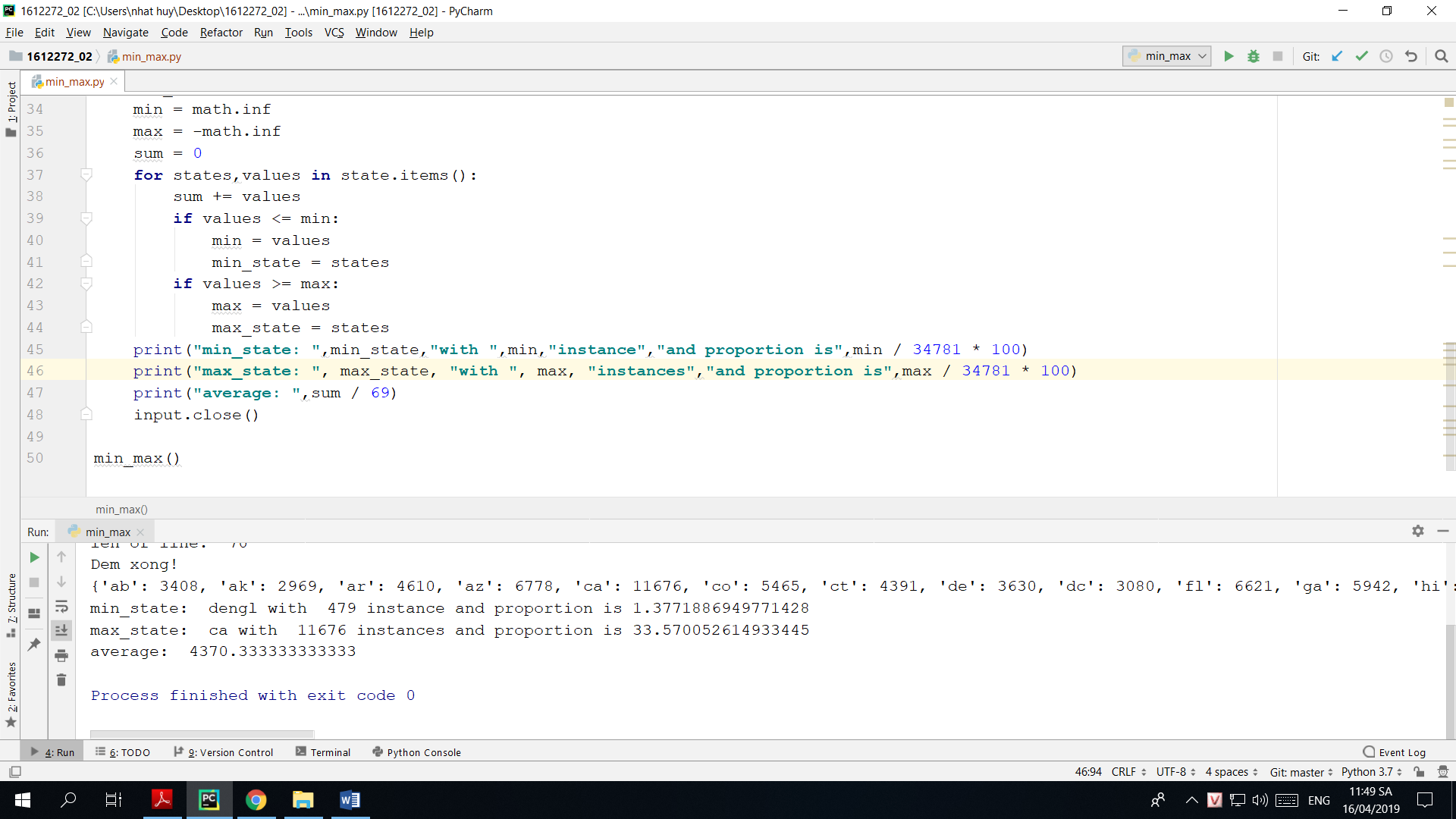


Figure 11: 6 vùng còn lại

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

1. **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

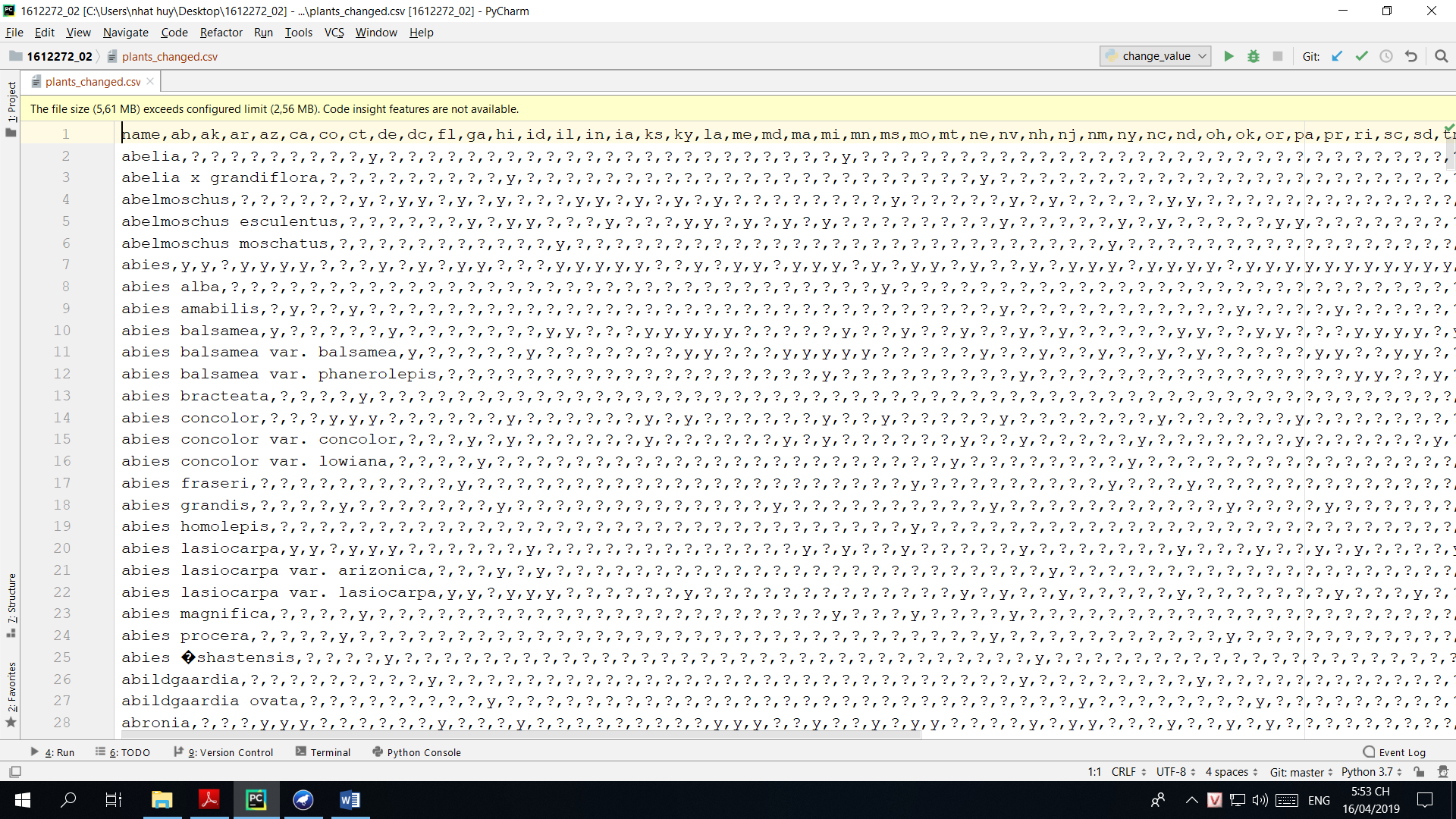


Figure 12: Nội dung file sau khi thay thế

Mở file plants\_changed.csv bằng Weka và tiến hành các bước tiếp theo

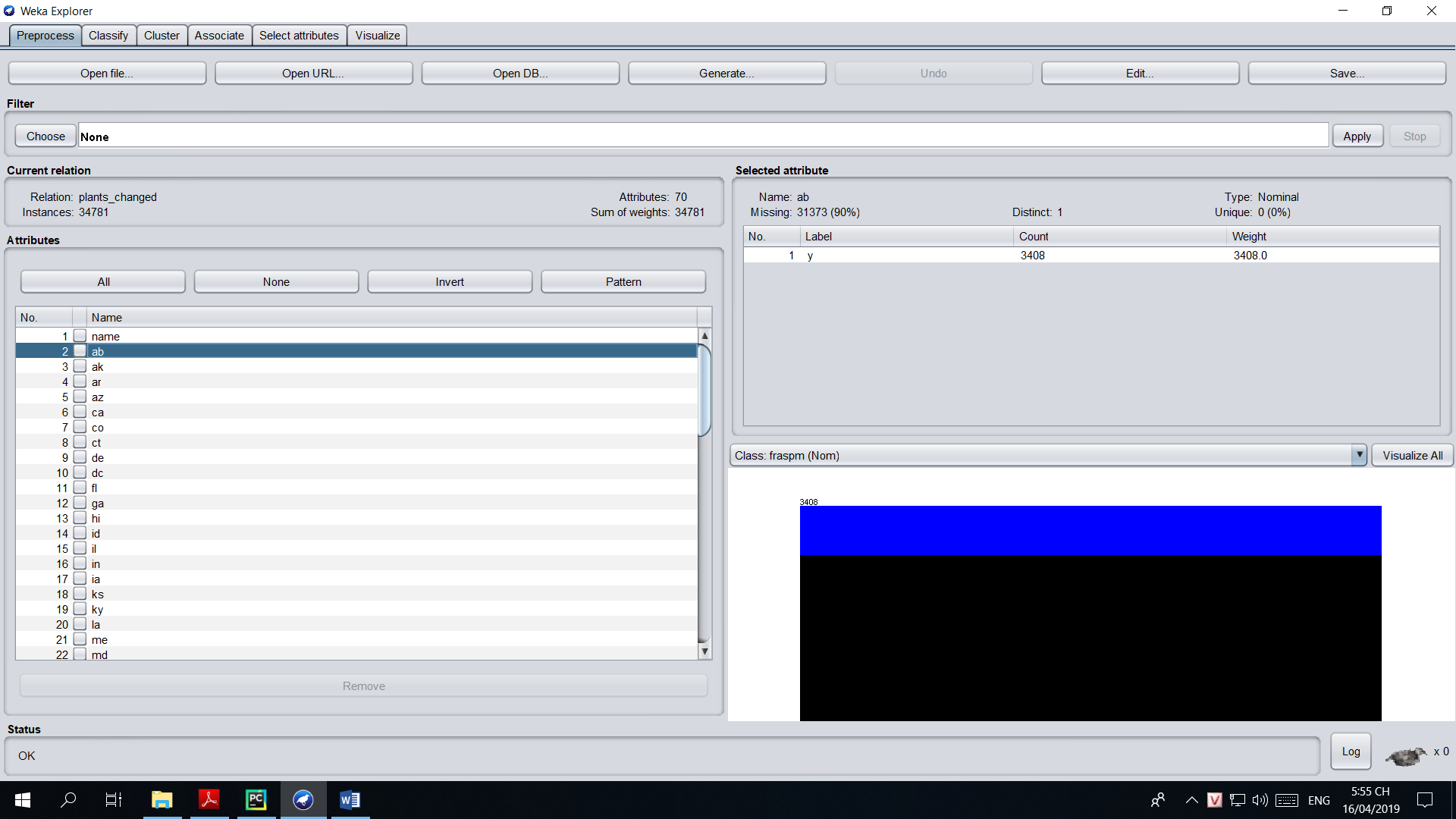


Figure 13: File plants\_changed.csv được mở bằng Weka

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

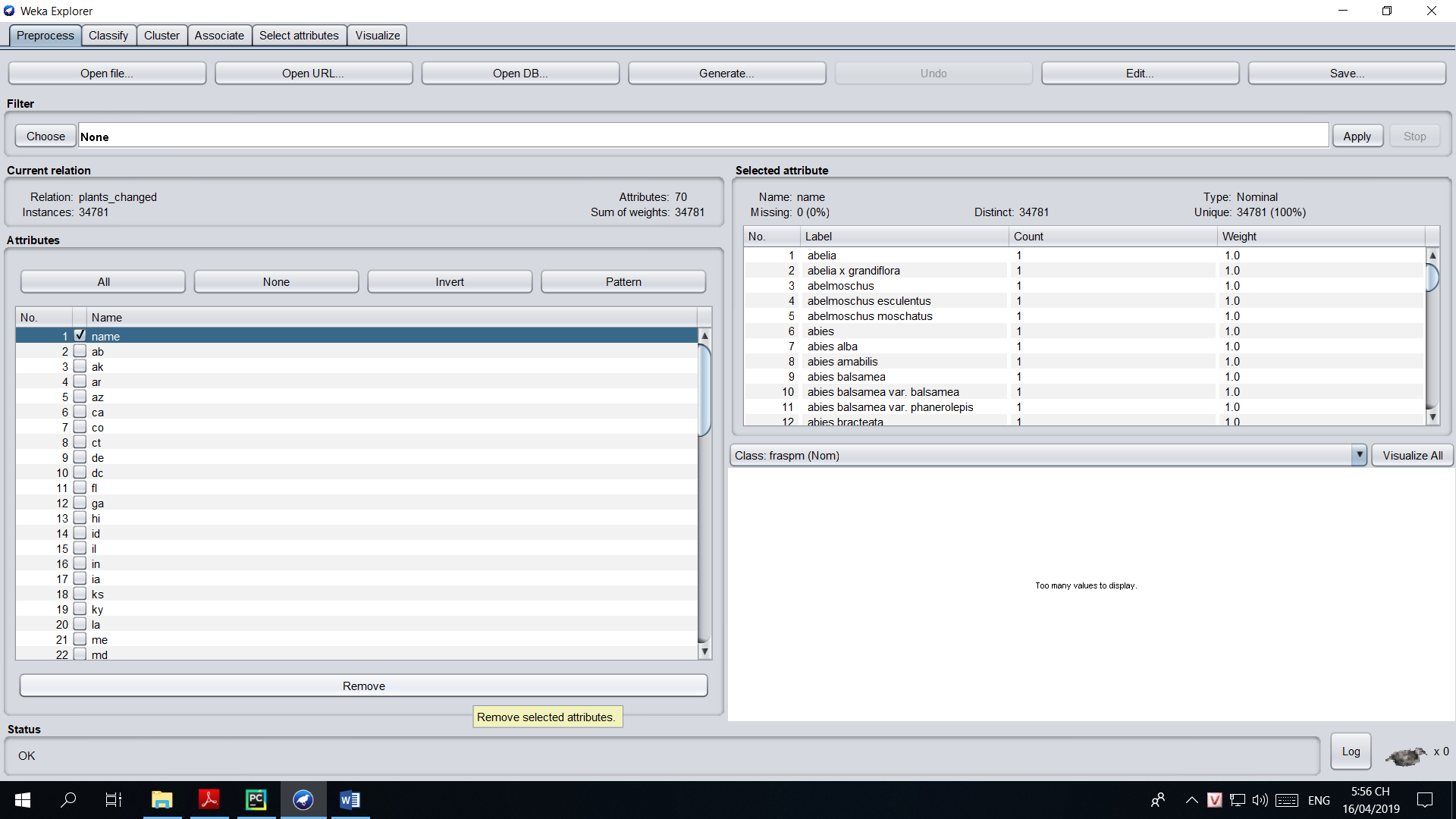


Figure 14: Chọn Remove để xóa thuộc tính name

Lưu file lại với tên là plants.arff

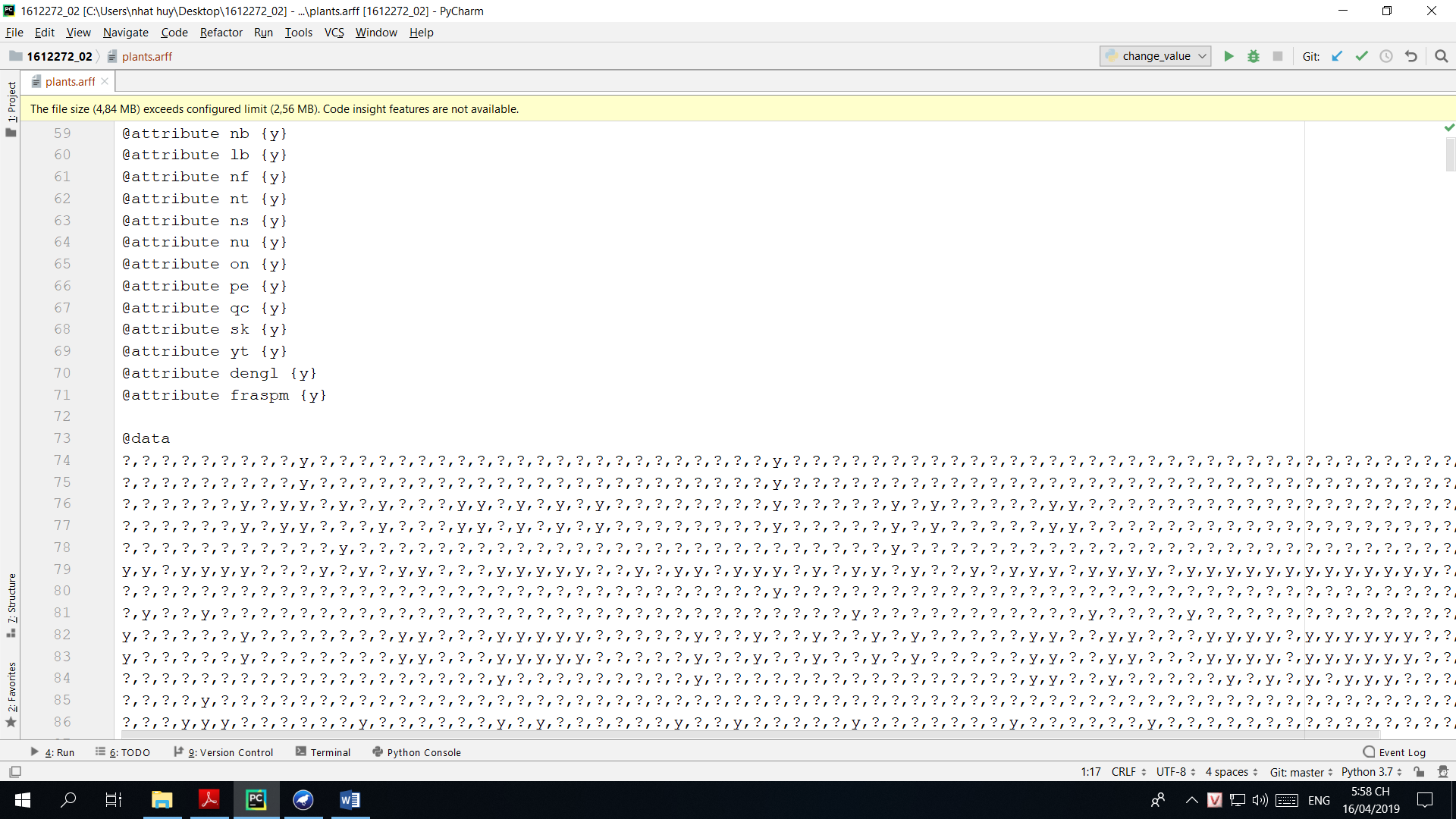


Figure 15: Nội dung file plants.arff

1. **CÂU 4**

Để khai thác tập phổ biến ta chuyển sang tab Associate

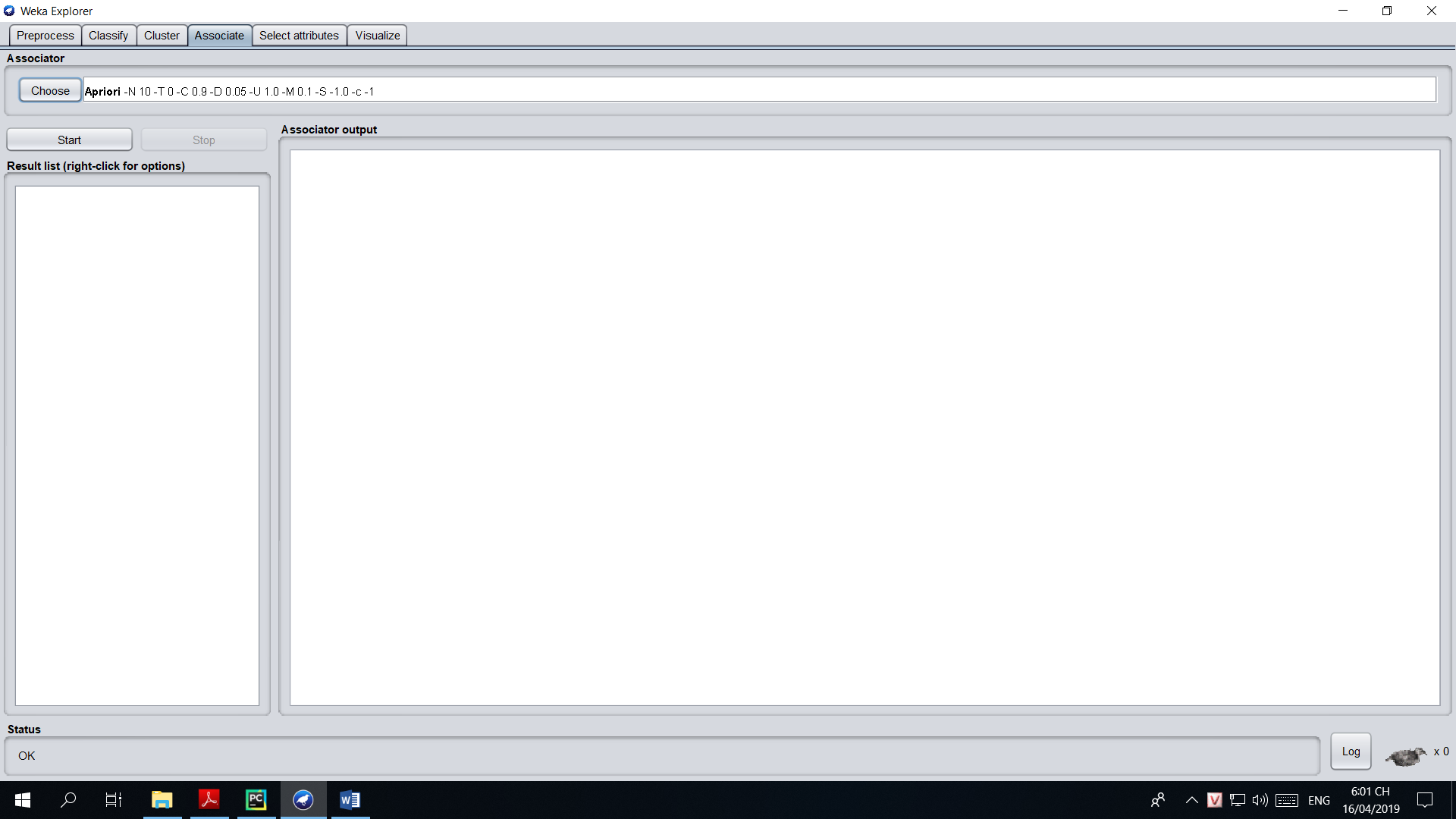


Figure 16: Tab Associate

Chọn thuật toán Apriori với các tham số sau

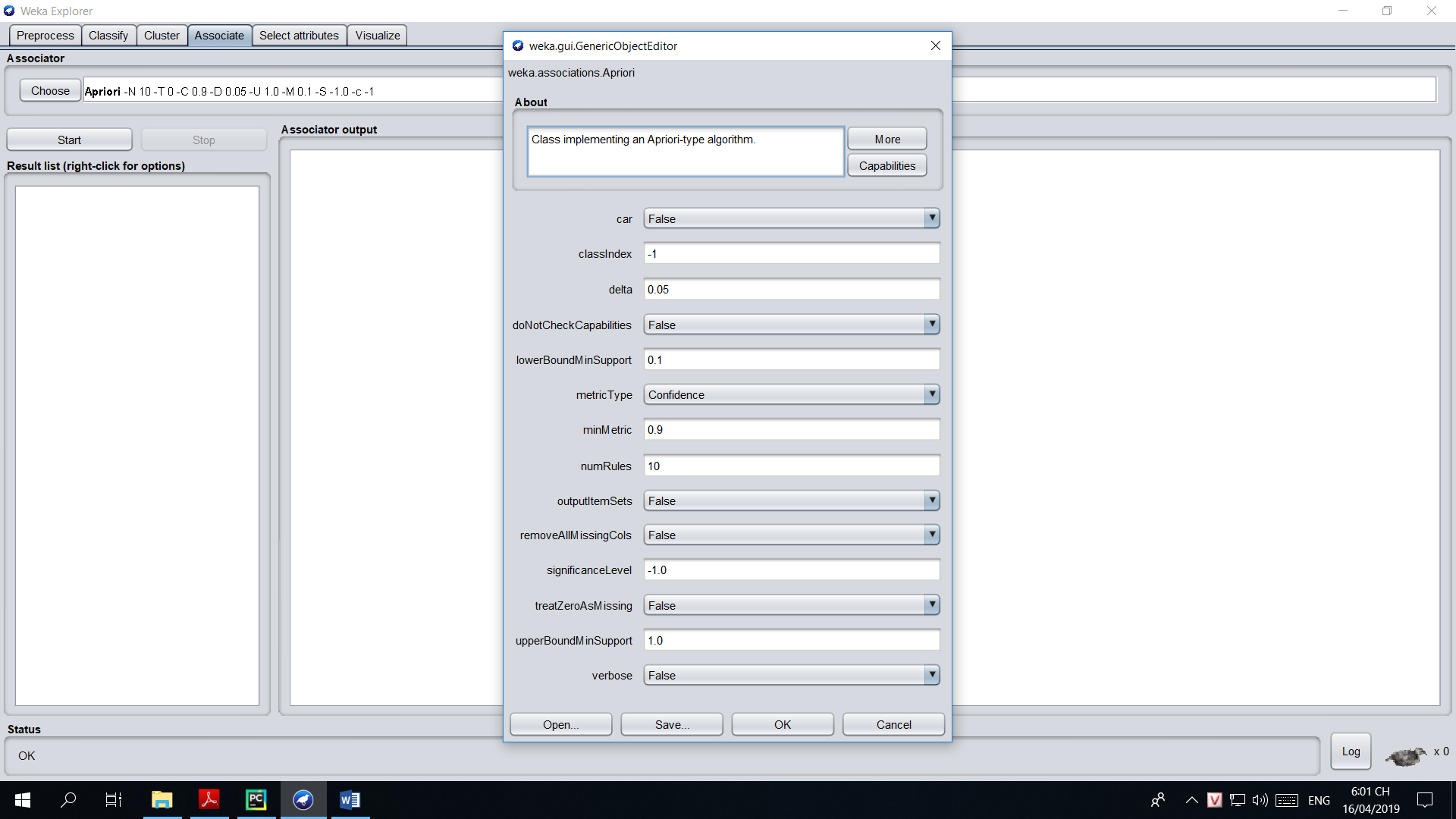


Figure 17: Thuật toán Apriori

Kết quả

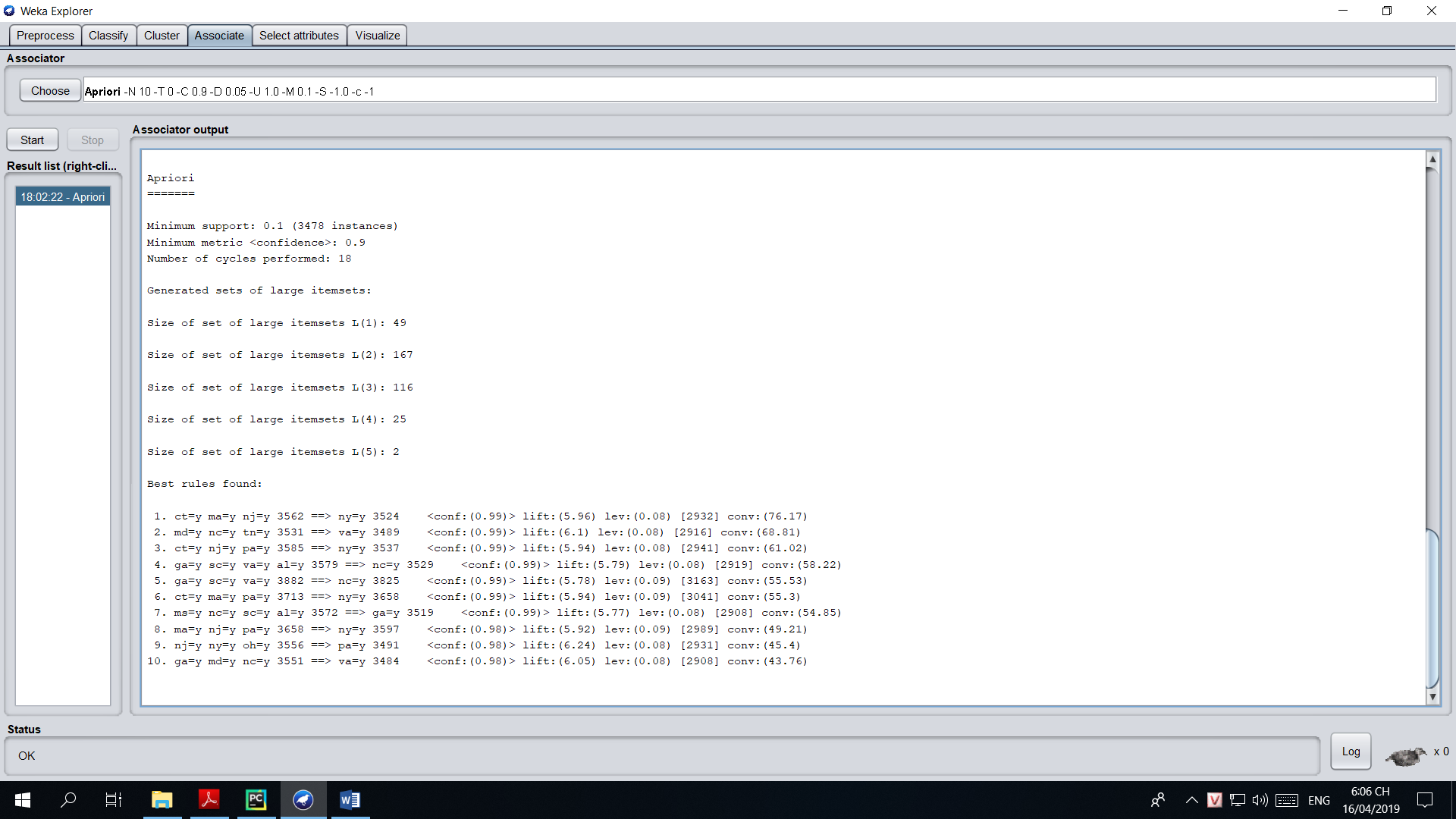


Figure 18: Kết quả chạy Apriori

|  |  |
| --- | --- |
| Kích thước | Số lượng |
| 1 hạng mục | 49 |
| 2 hạng mục | 167 |
| 3 hạng mục | 116 |
| 4 hạng mục | 25 |
| 5 hạng mục | 2 |

1. **CÂU 5**

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

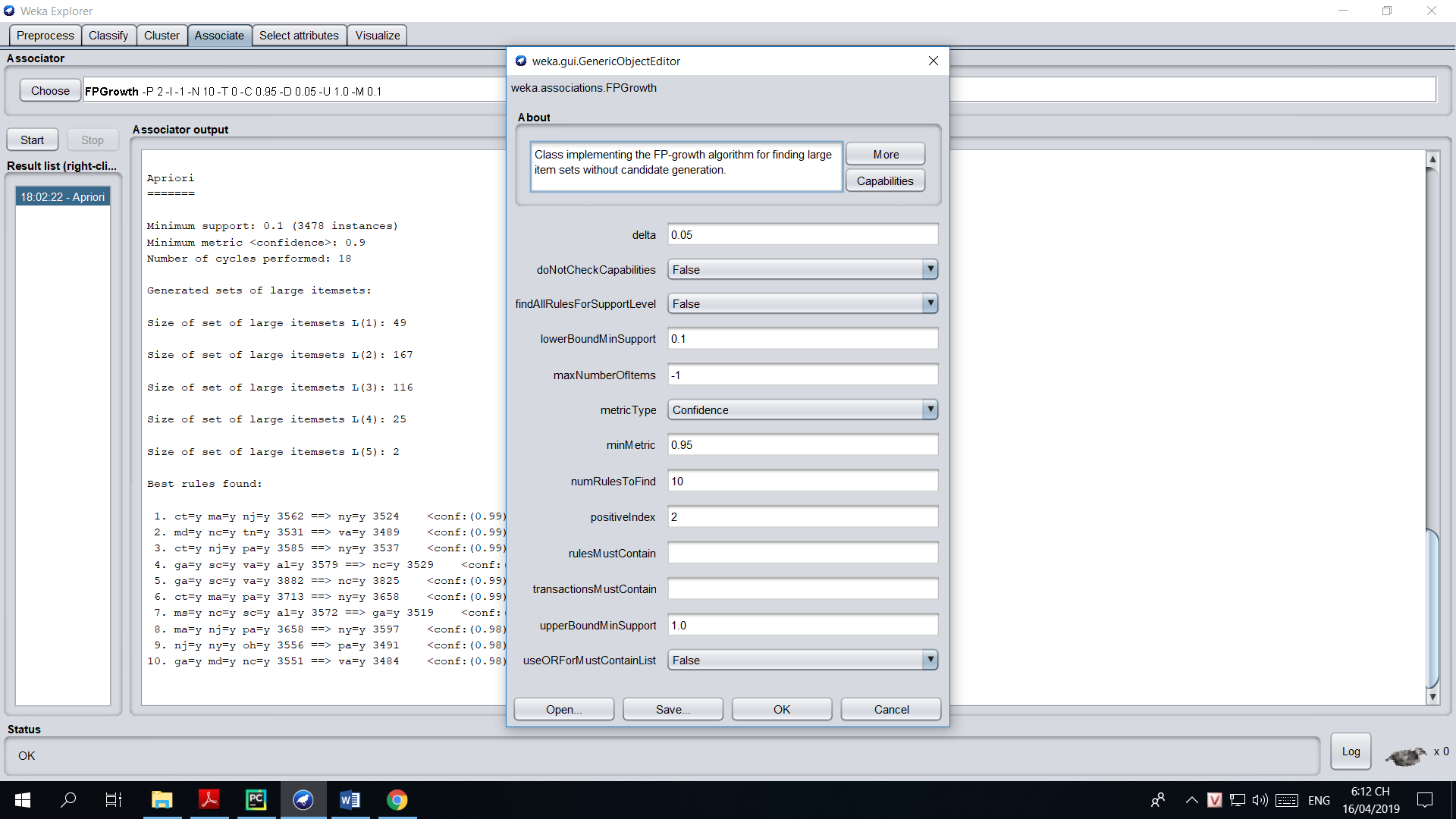


Figure 19: Tham số thuật toán FP-Growth

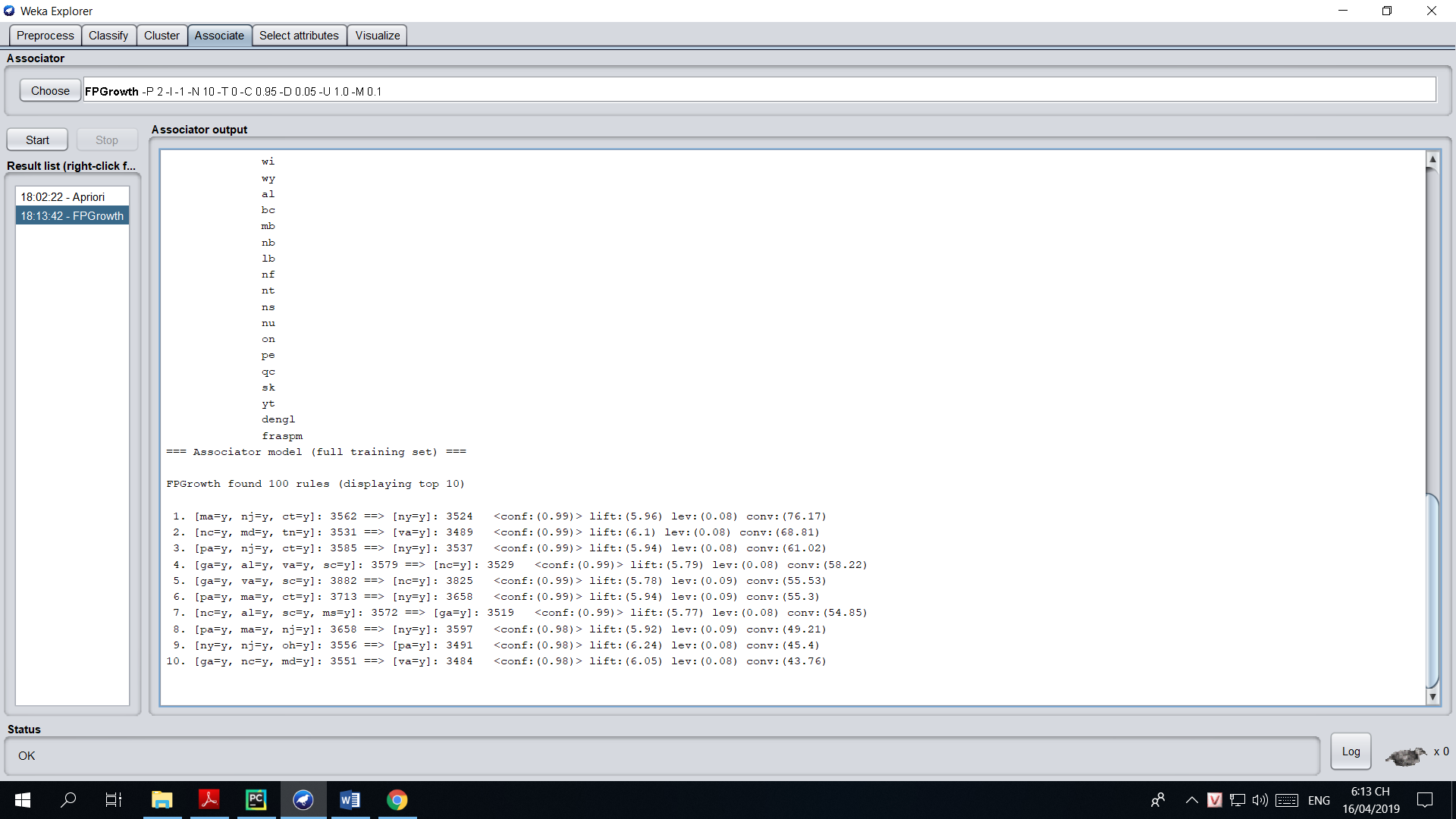


Figure 20: Kết quả chạy FP-Growth

|  |  |
| --- | --- |
| 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 |