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

**🙥🟊🙧**

**BÀI TẬP 1: 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** |
| 1 | a. Định nghĩa sự hợp nhất dữ liệu | 100% |  |
| b. Có vấn đề nhận hiện thực thể hay không? Giải quyết (nếu có) | 100% |  |
| c. Có vấn đề dữ liệu dư thừa không? Giải quyết (nếu có) | 100% |  |
| d. Có mâu thuẫn dữ liệu không? Giải quyết (nếu có) | 100% |  |
| e. Tích hợp 2 dataset | 100% |  |
| f. Chụp lại màn hình | 100% |  |
| 2 | a. Xem thuộc tính age và trả lời các câu hỏi | 100% |  |
| b. Liệt kê five-number summary của thuộc tính age | 100% |  |
| c. Bao nhiêu thuộc tính số, có thứ tự, rời rạc/danh sách? | 100% |  |
| d. Giải thích đồ thị trong Explorer | 100% |  |
| e. Dán ảnh chụp các đồ thị vào bài làm | 100% |  |
| f. Nhận xét về đồ thị | 100% |  |
| g. Dán đồ thị bạn cho rằng có khả năng đoán bệnh tim tốt nhất | 100% |  |
| h. Những cặp thuộc tính nào tương quan? | 100% |  |
| 3 | a. Có bao nhiêu thuộc tính trong dataset? | 100% |  |
| b. Liệt kê các phương pháp lọc thuộc tính | 100% |  |
| c. So sánh các phương pháp trong textbook và weka | 100% |  |
| 4 | a. Dữ liệu thiếu | 80% | Chưa cài đặt một phương pháp trong weka |
| b. Dữ liệu nhiễu | 100% |  |
| c. Dữ liệu tạp | 100% |  |
| d. Lưu dataset đã làm sạch | 100% |  |
| 5 | a. Xây dựng thuộc tính | 100% |  |
| b. Chuẩn hóa | 100% |  |
| c. Chọn 1 phương pháp chuẩn hóa | 100% |  |
| d. Lưu dataset đã chuẩn hóa | 100% |  |
| 6 | Lấy mẫu | 100% |  |

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

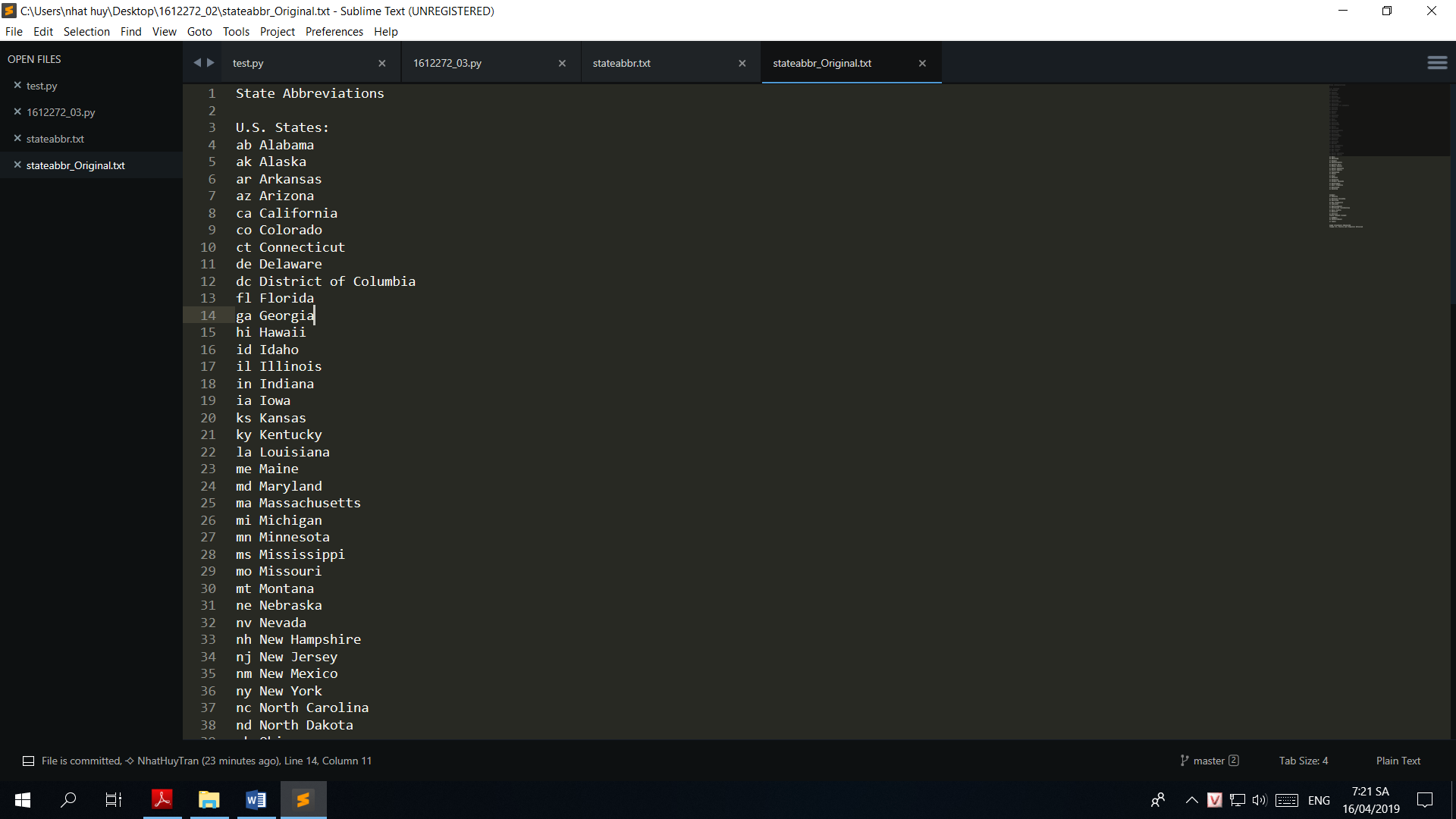


Figure 1: Nội dung file stateabbr.txt

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

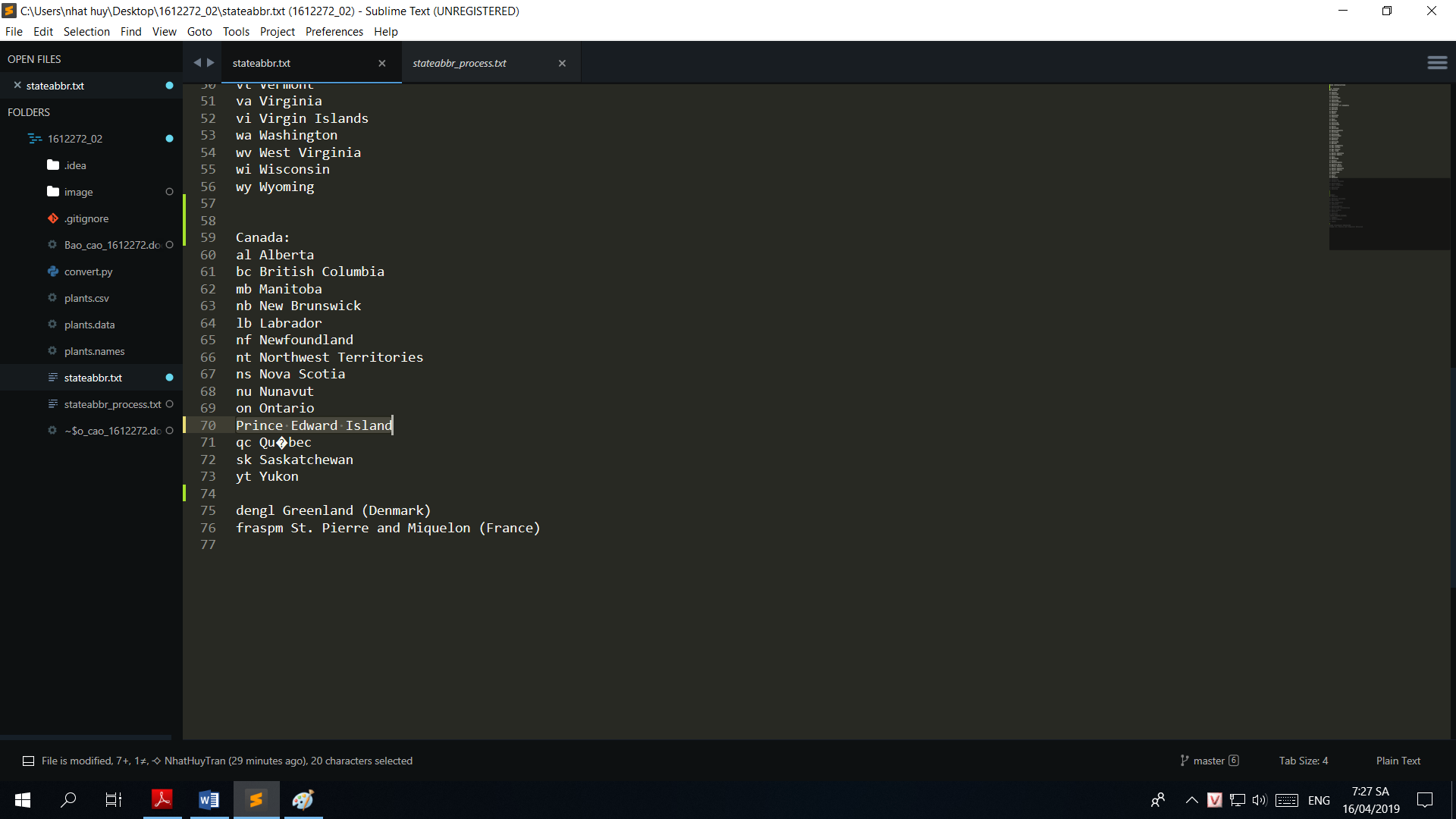


Figure 2: Thành phố Prince Edward Island trong stateabbr.txt

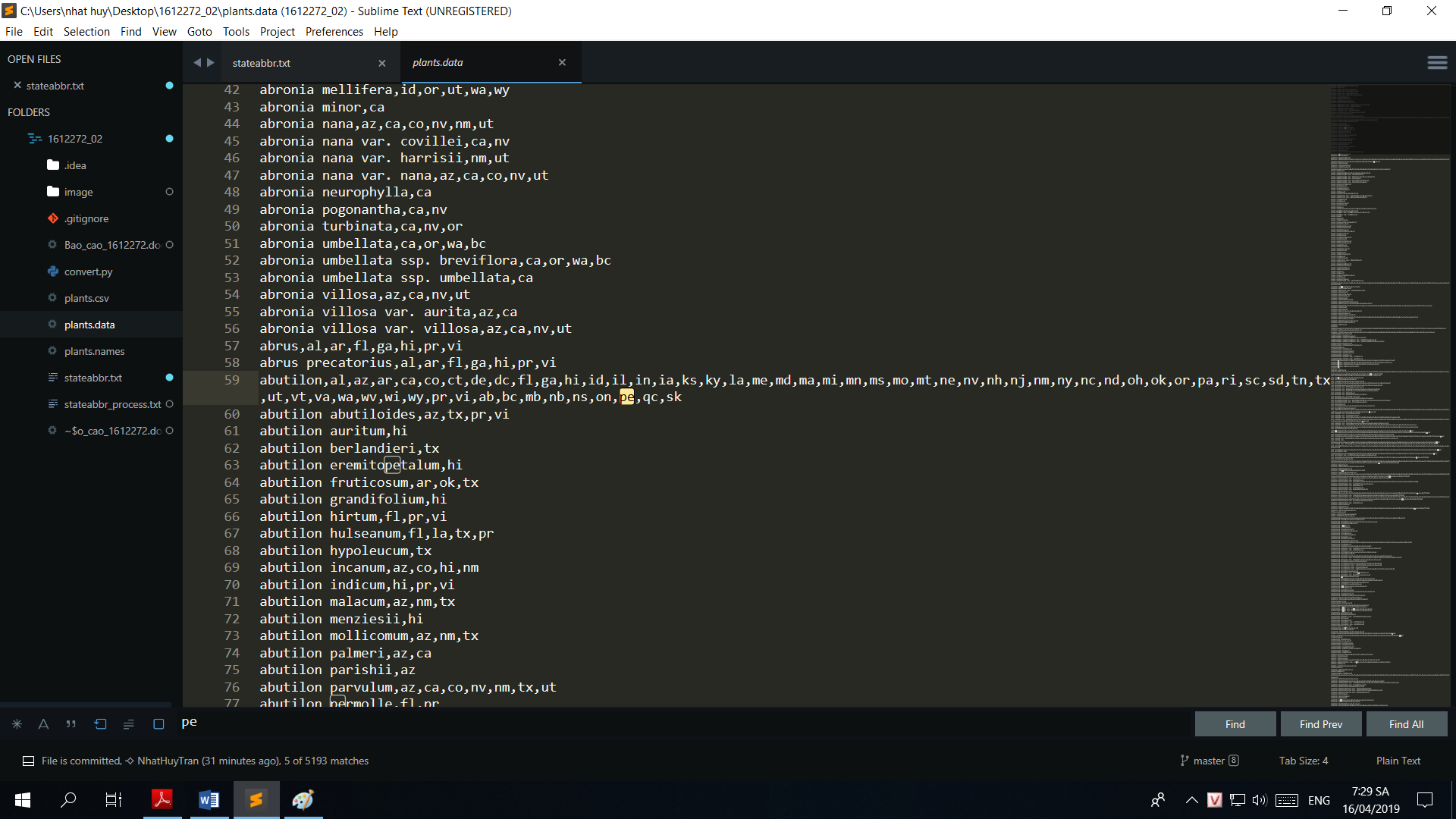


Figure 3: Ký hiệu pe được sử dụng trong plants.data

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

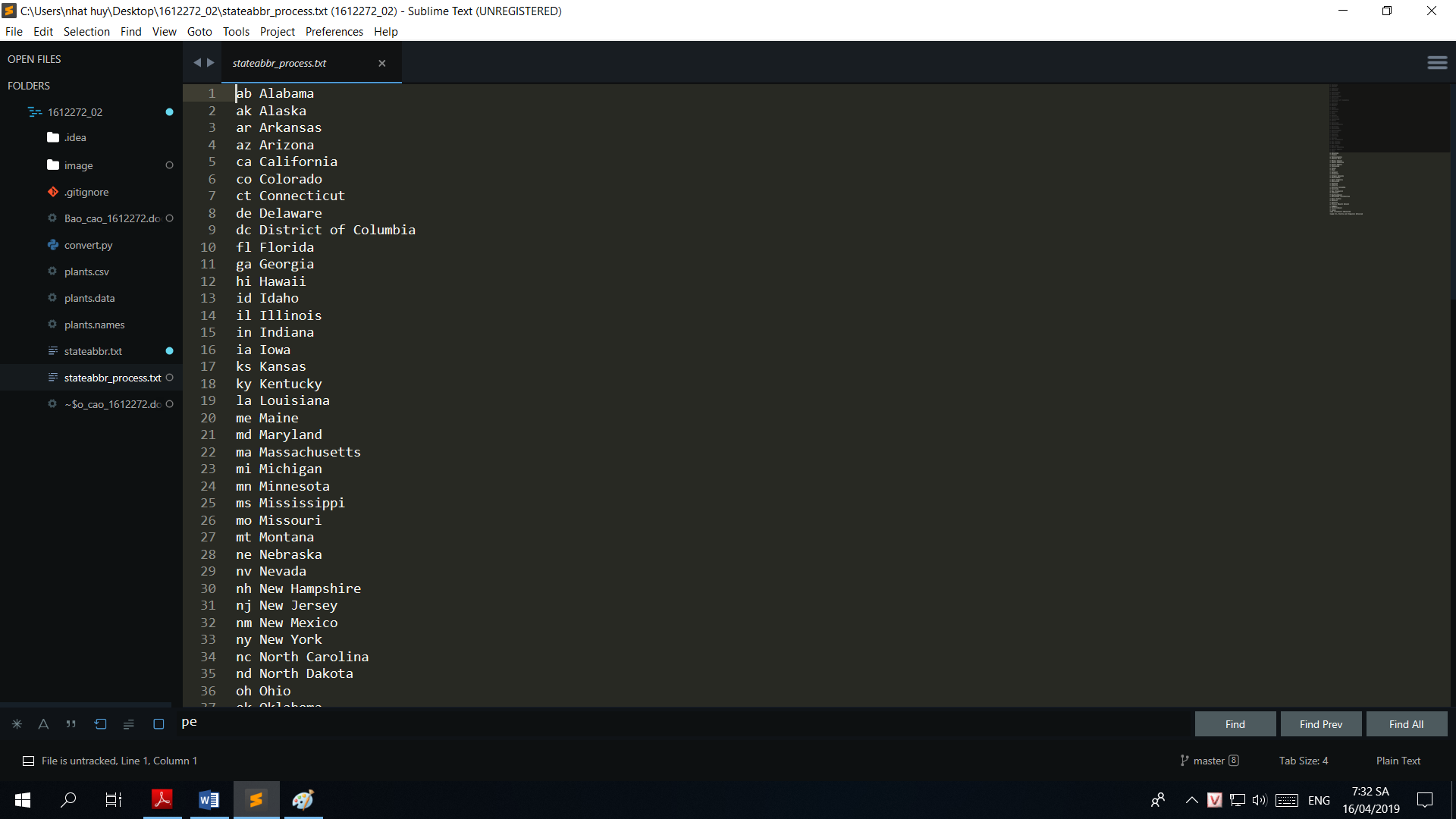


Figure 4: File stateabbr\_process.txt

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

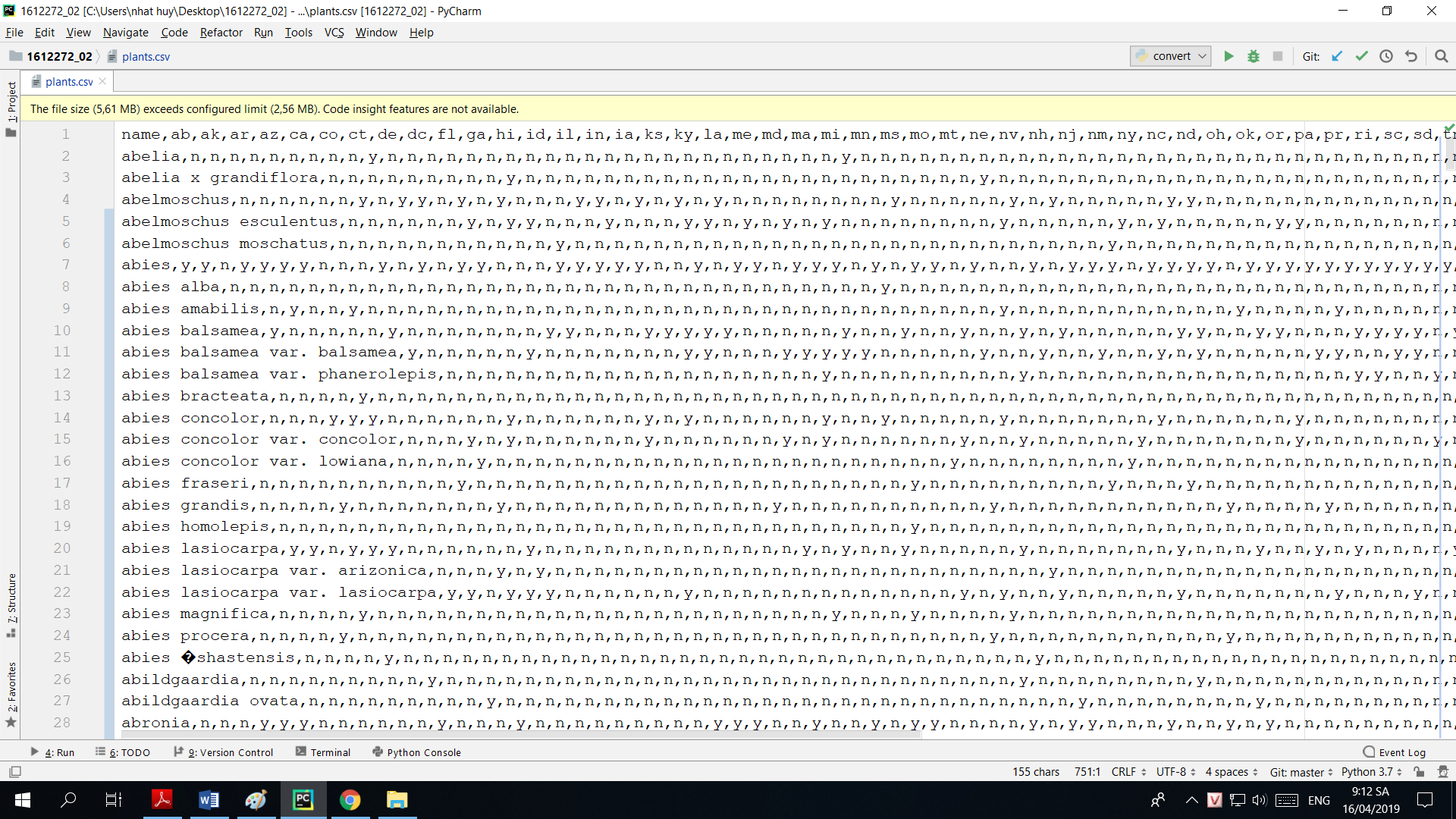


Figure 5: File plants.data sau khi chuyển

1. **CÂU 2**

Ta tiến hành mở file plants.csv bằng Weka

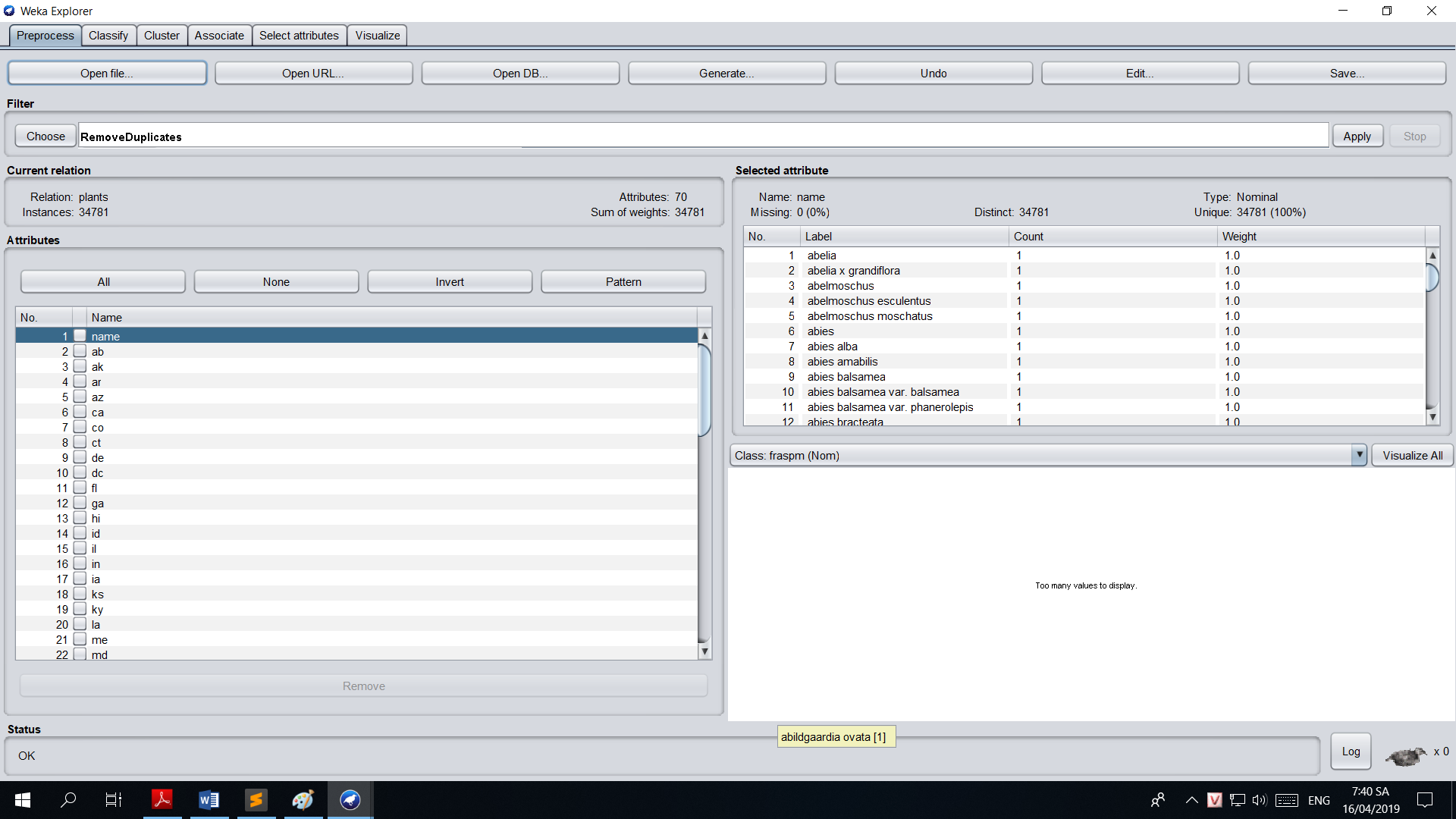


Figure 6: File plants.csv mở bằng Weka

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 7: 31 vùng đầu tiên

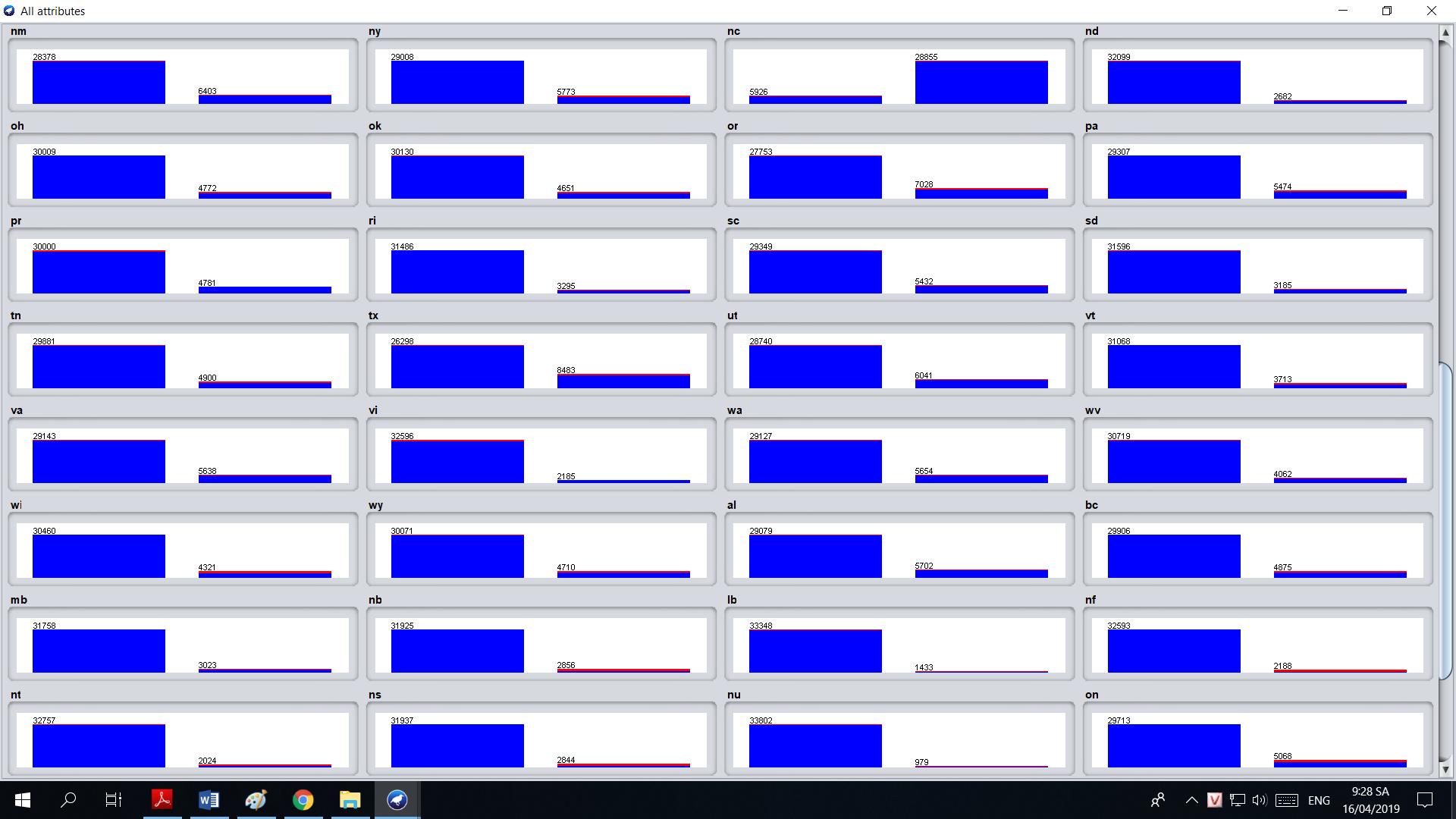


Figure 8: 32 vùng tiếp theo

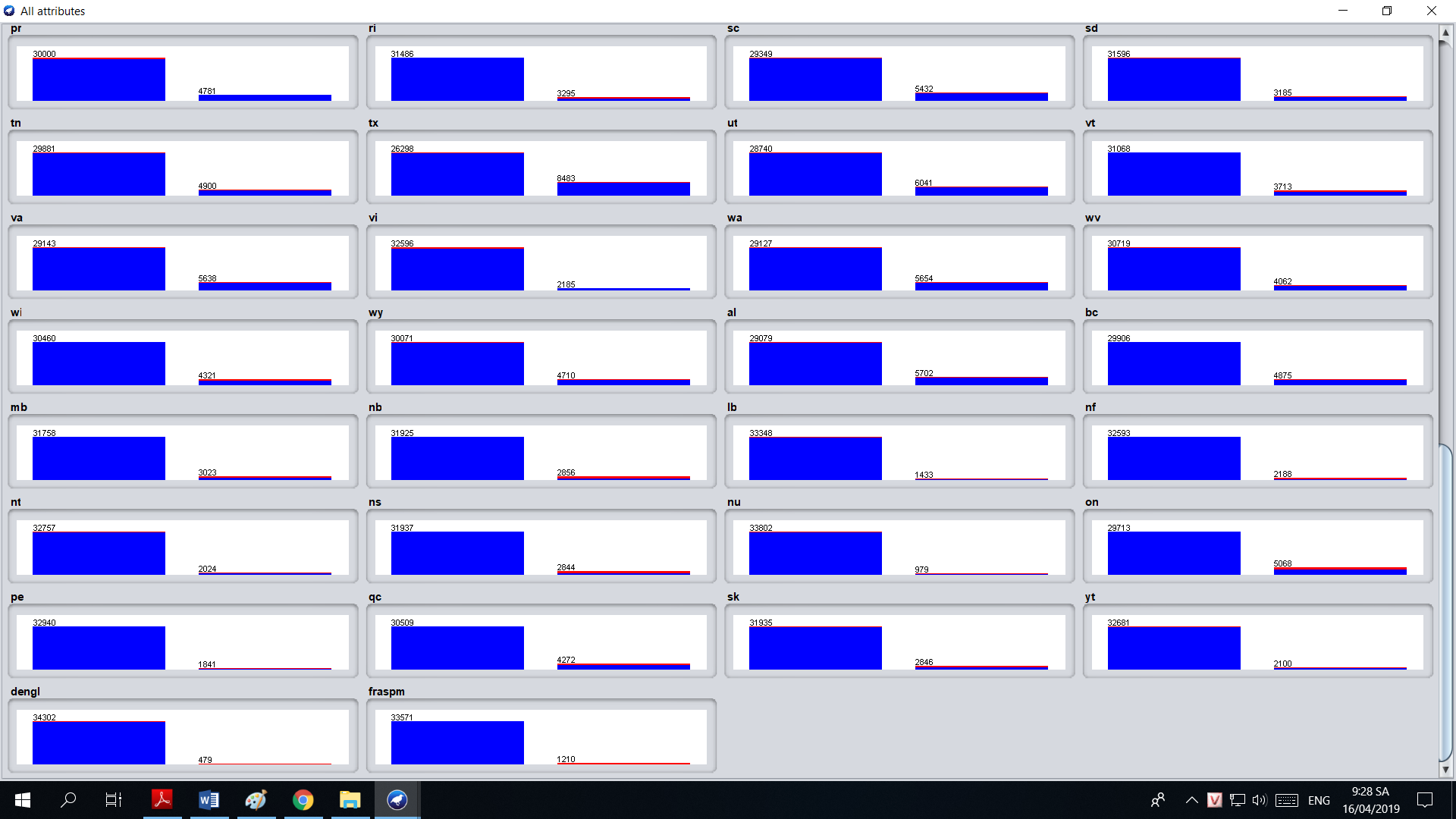
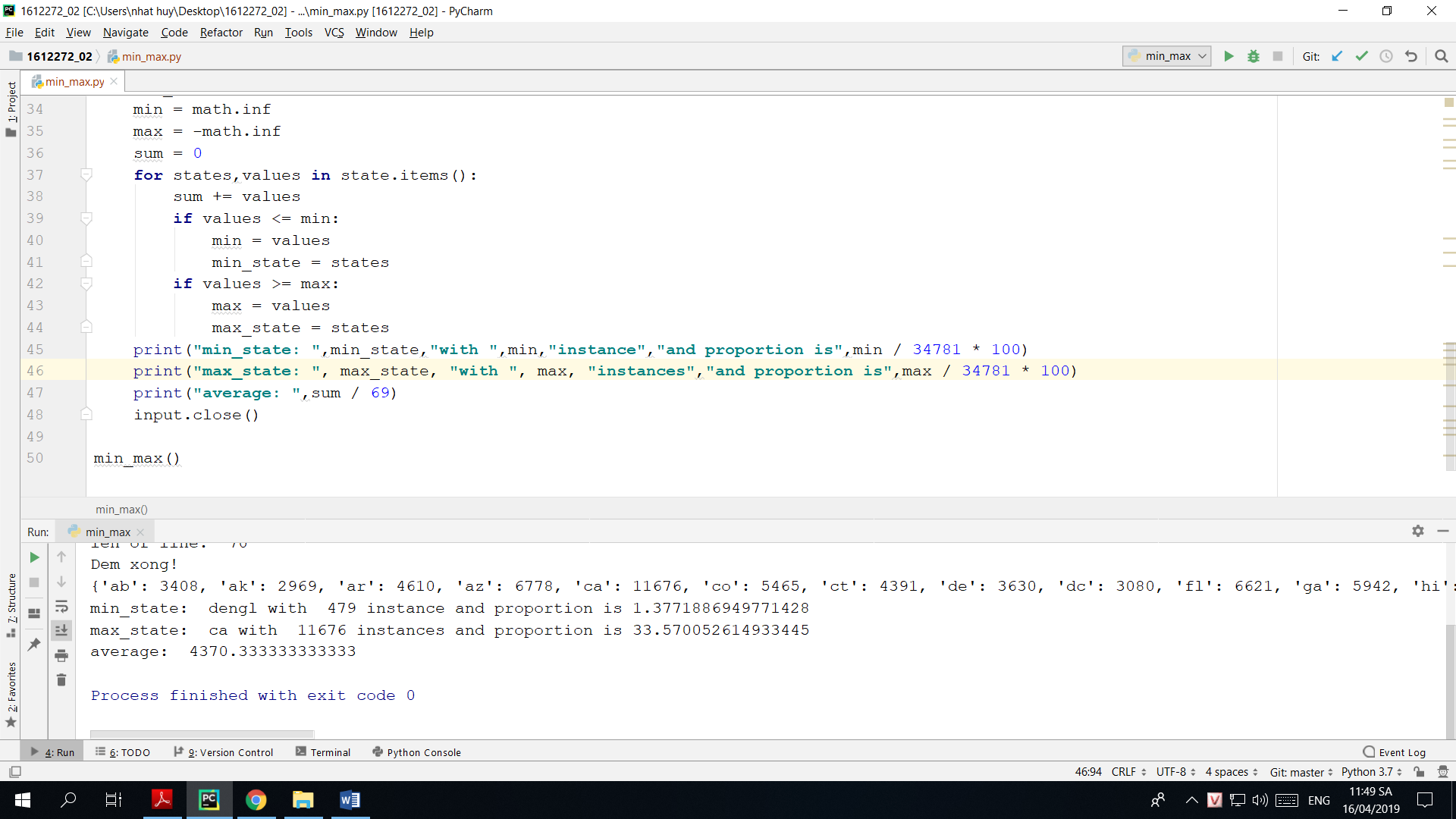


Figure 9: 6 vùng còn lại

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

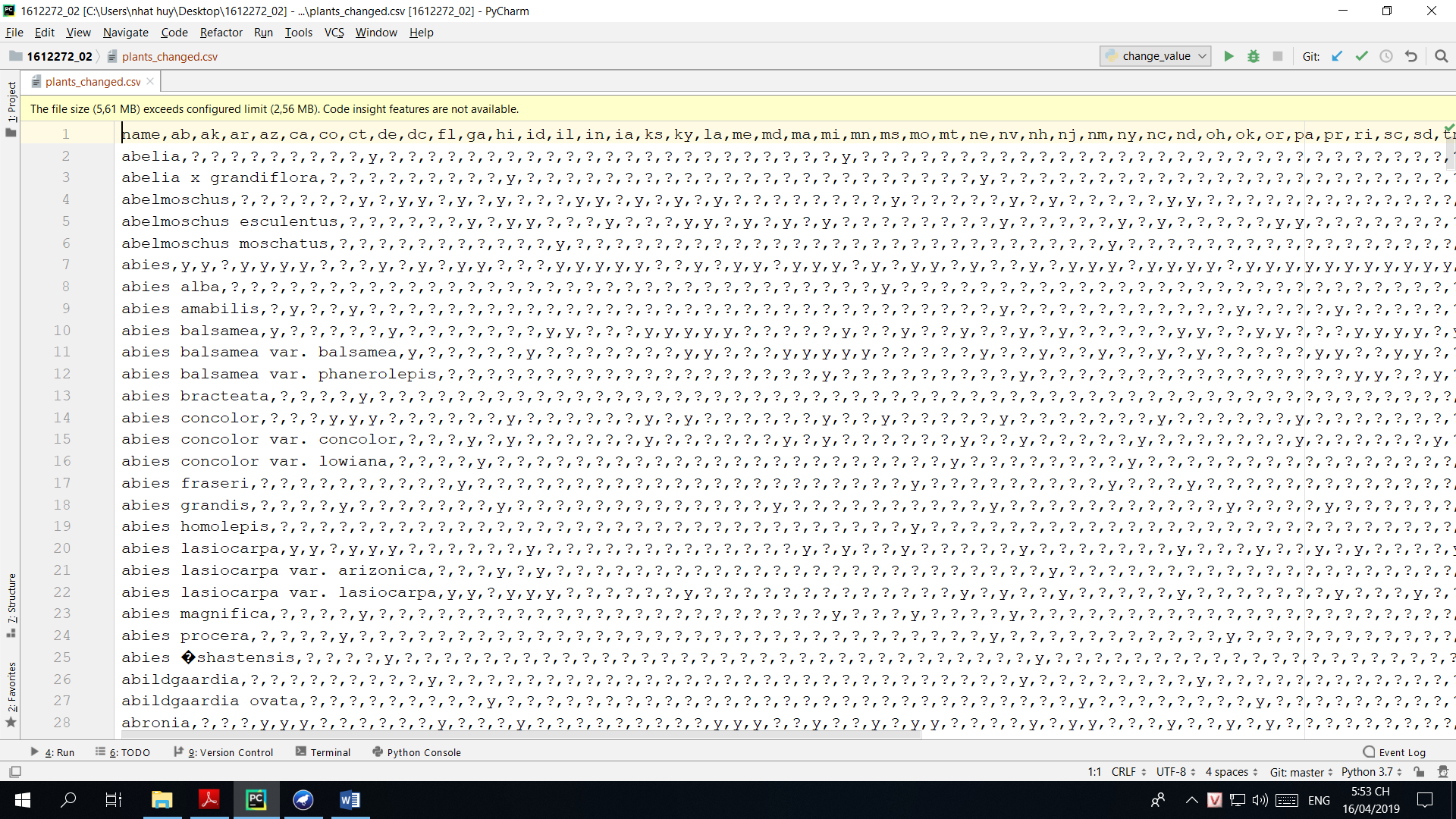


Figure 10: Nội dung file sau khi thay thế

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

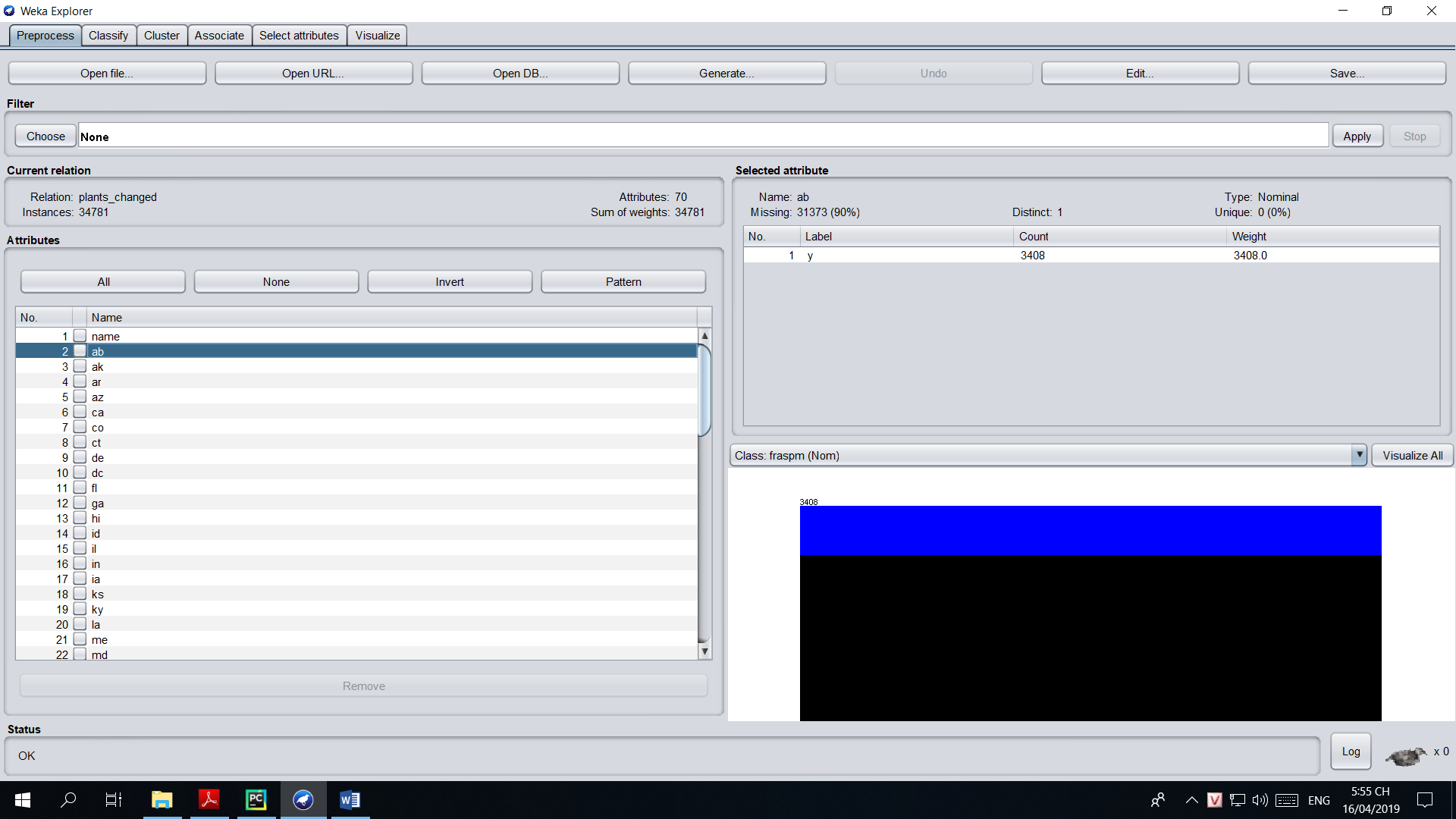


Figure 11: File plants\_changed.csv được mở bằng Weka

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

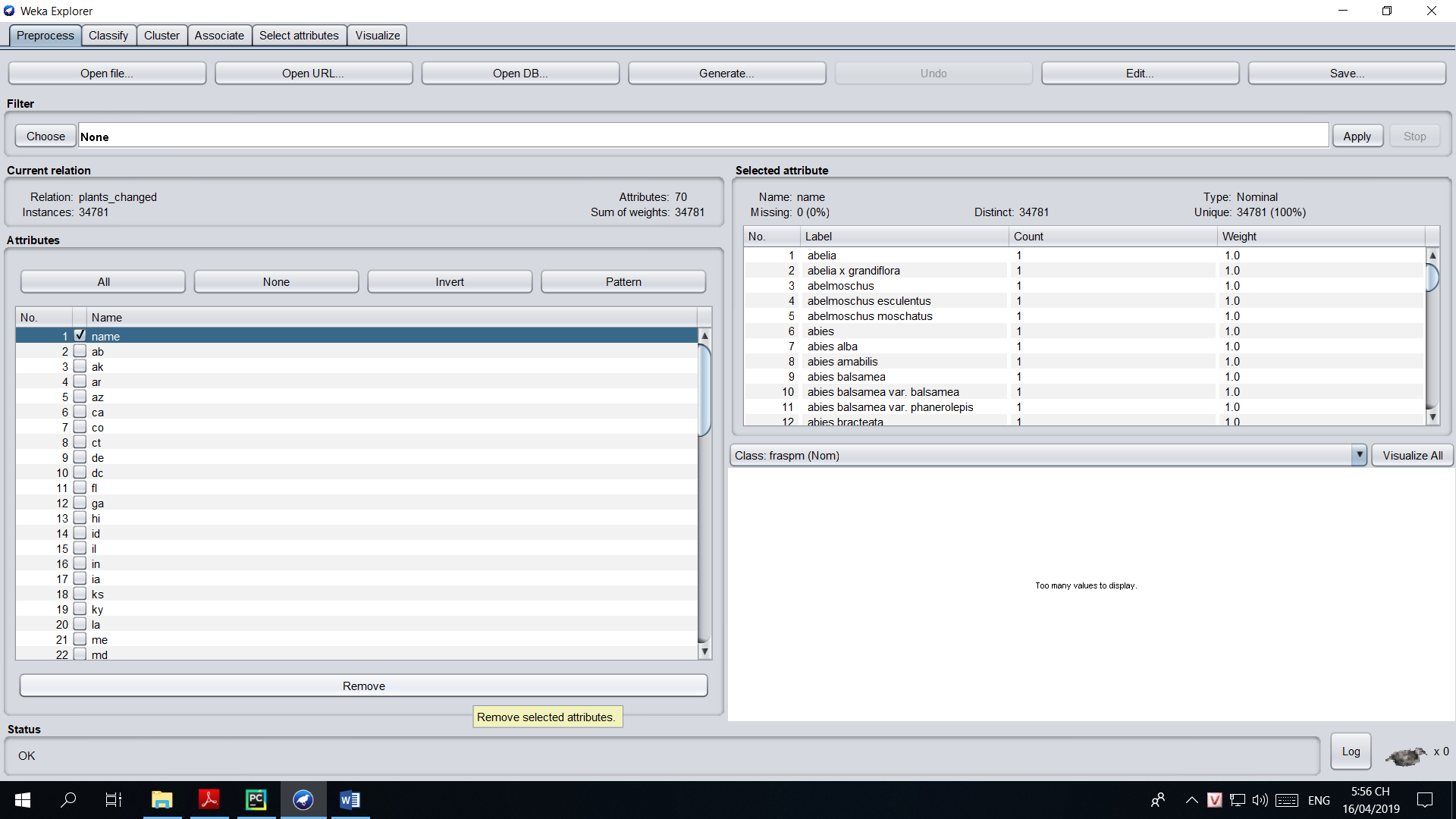


Figure 12: Chọn Remove để xóa thuộc tính name

Lưu file lại với tên là plants.arff

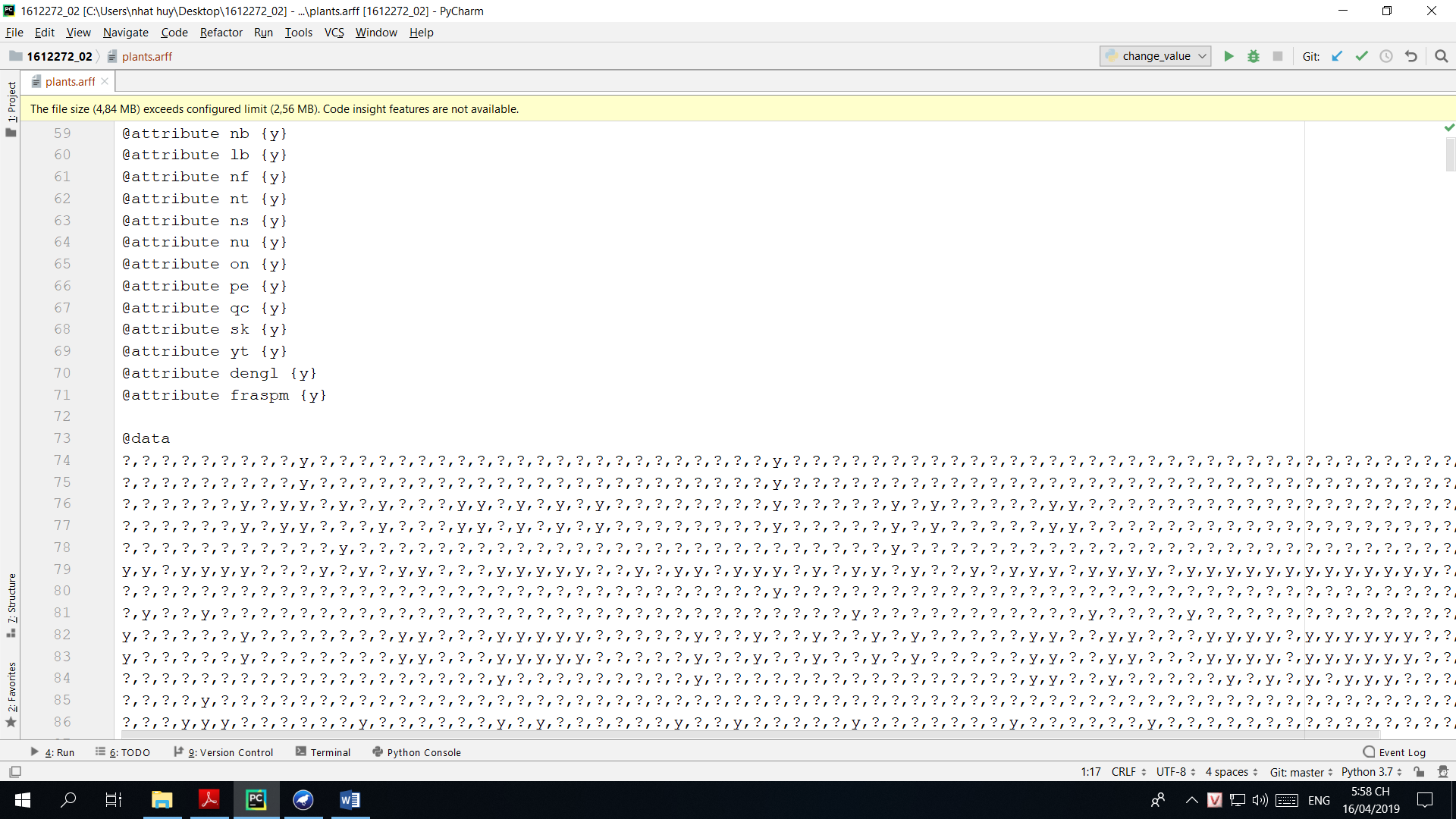


Figure 13: Nội dung file plants.arff

1. **CÂU 4**

Để khai thác tập phổ biến ta chuyển sang tab Associate

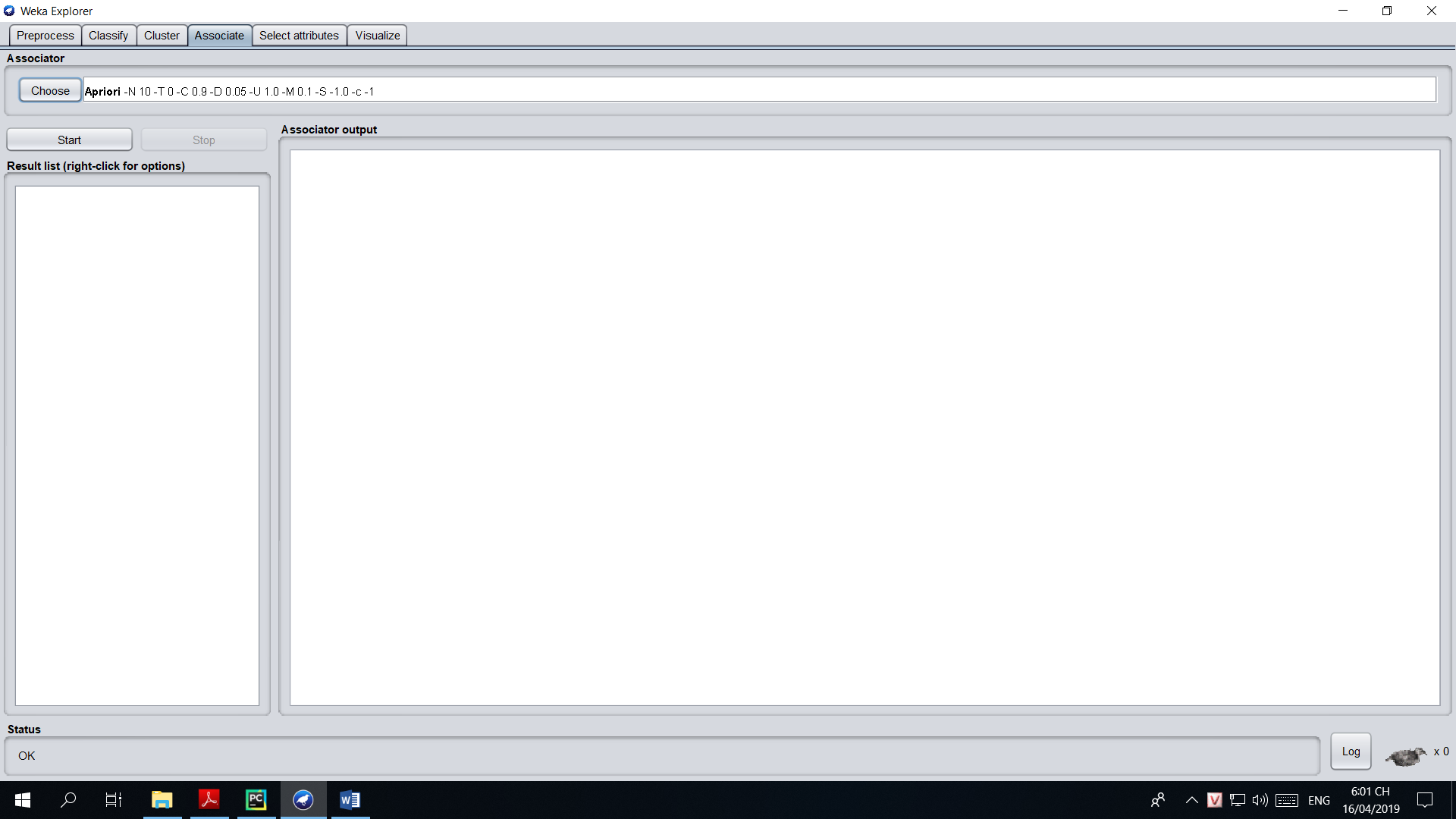


Figure 14: Tab Associate

Chọn thuật toán Apriori với các tham số sau

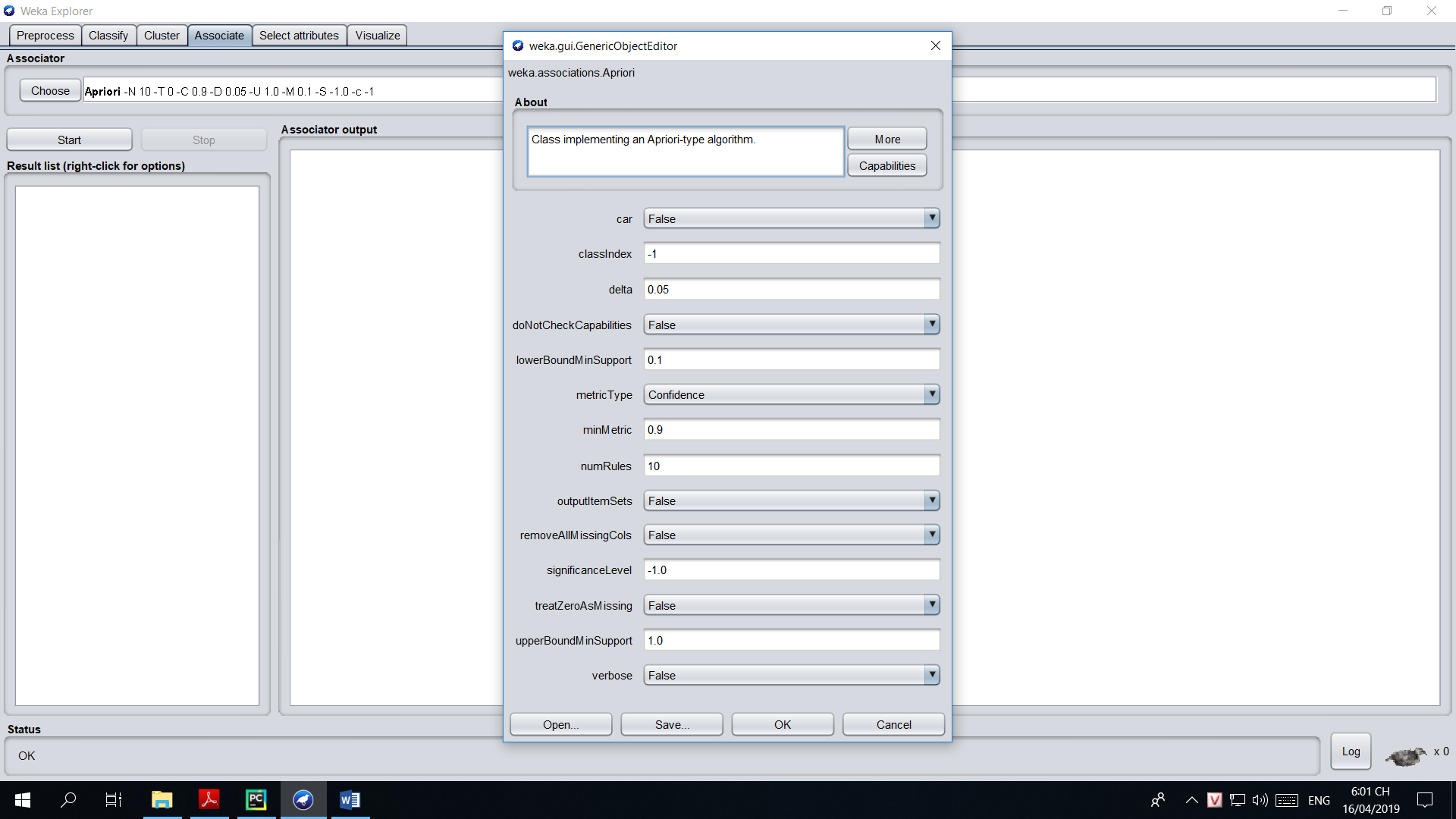


Figure 15: Thuật toán Apriori

Kết quả

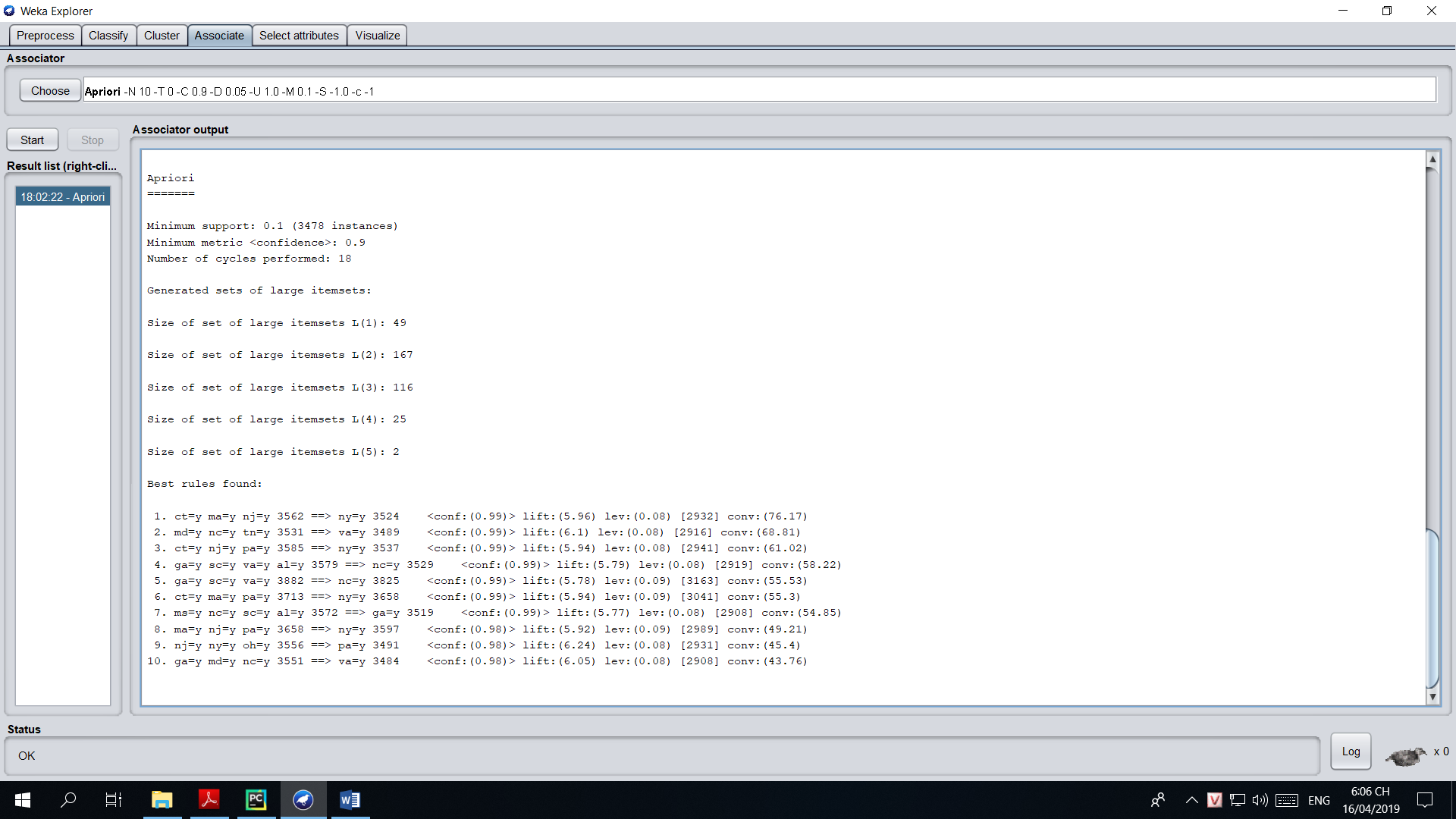


Figure 16: Kết quả chạy Apriori

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

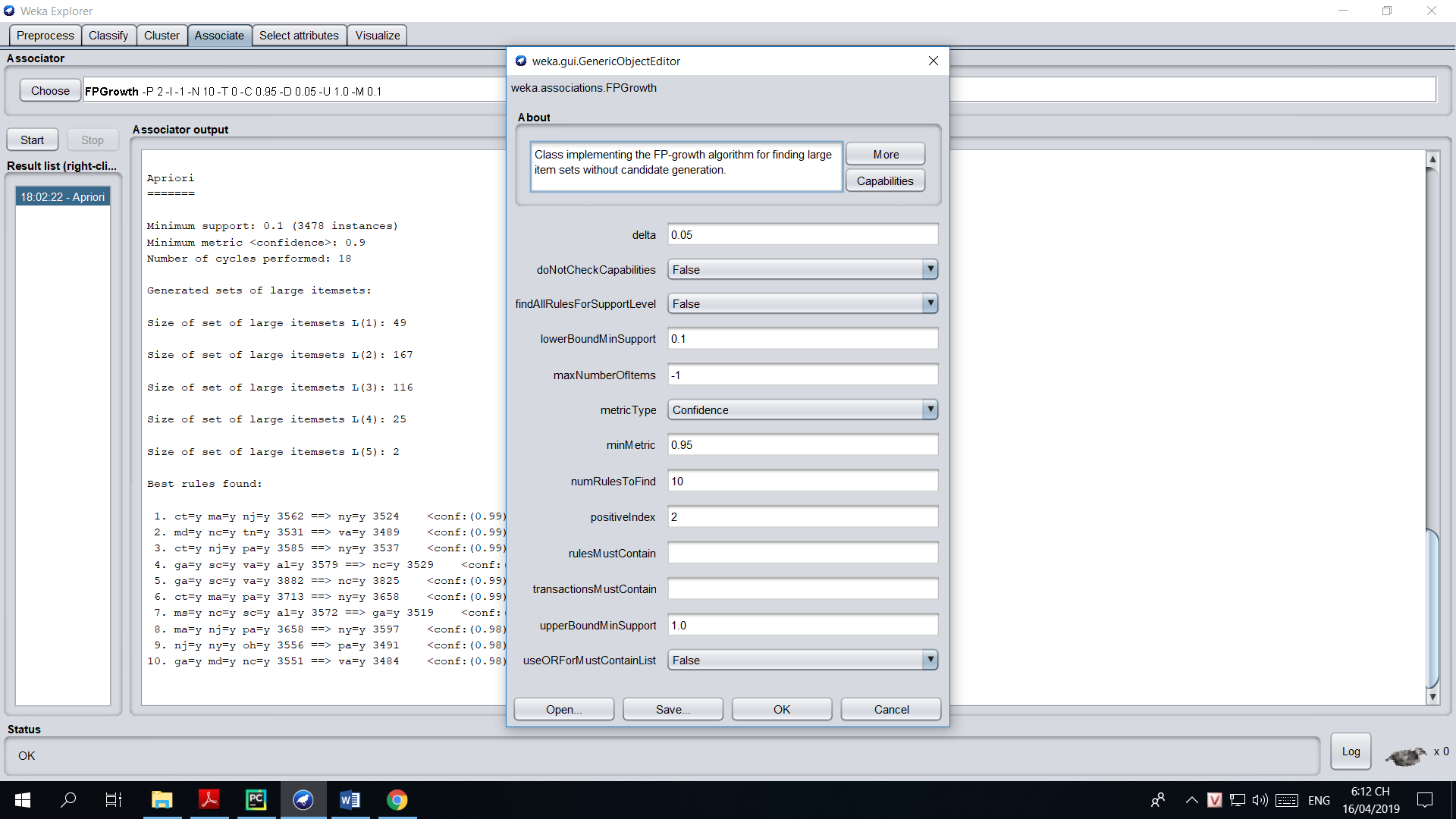


Figure 17: Tham số thuật toán FP-Growth

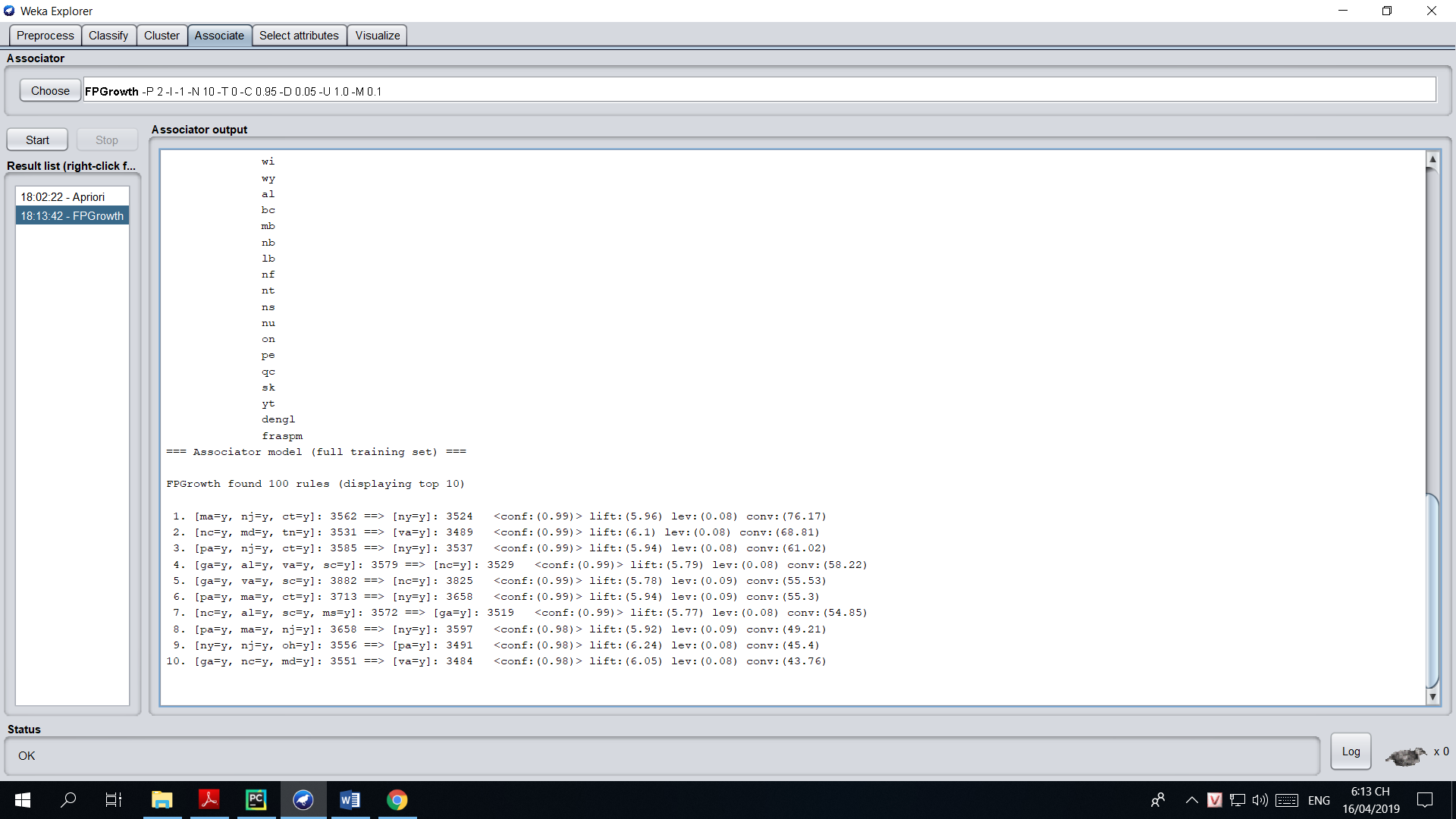


Figure 18: Kết quả chạy FP-Growth

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