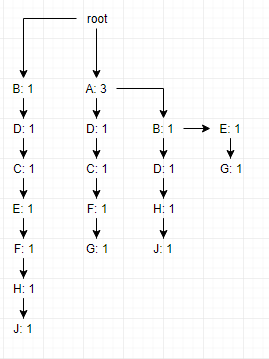
1. Lý thuyết:
2. Tìm hiểu và trình bày chi tiết một phương pháp cải tiến quá trình tìm luật kết hợp từ tập phổ biến. Giải thích vì sao nó hiệu quả
3. Cho CSDL sau và minsupp = 50%, minconf = 100%

|  |  |
| --- | --- |
| **TID** | **Items\_bought** |
| **100** | **I, B, F, D, E, C, H, J** |
| **200** | **F, C, F, G, A, D, C** |
| **300** | **B, J, D, A, H** |
| **400** | **E, A, B, E, G** |

1. - Sử dụng thuật toán Apriori để tìm tất cả các tập phổ biến:

* {A, B, D, C, E, F, G, H, J, AB, AD, AG, BD, BH, BE, BJ, CD, CF, DF, DH, DJ, HJ, BDH, BDJ, BHJ, CDF, DHJ, BDHJ}.
* Các tập Large k-item phát sinh:
* Sử dụng thuật toán FP-Growth để tìm tất cả các tập phổ biến:
* Cây FP:



(Hình A2c)

* Bảng xây dựng cây cơ sở mẫu điều kiện và cây FP điều kiện, mẫu phổ biến:

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Cơ sở mẫu điều kiện** | **FP-Tree điều kiện** | **Các mẫu phổ biến** |
| J | {BDCEFH: 1, ABDH: 1} | {B: 2, D: 2, H: 2} - J | J, BJ, DJ, HJ, BDJ, BHJ, DHJ, BDHJ |
| H | {BDCEF: 1, ABD: 1} | {B: 2, D: 2} - H | H, BH, DH, BDH |
| G | {ADCF: 1, ABE: 1} | {A: 2} - G | G, AG |
| F | {BDCE: 1, ADC: 1} | {C: 2, D: 2} - F | F, CF, DF, CDF |
| E | {BDC: 1, AB: 1} | {B: 2} - E | E, BE |
| C | {BD: 1, AD: 1} | {D: 2} - C | C, CD |
| D | {B: 1, A: 1, AB: 1} | {B: 2, A: 2} - D | D, BD, AD |
| B | {A: 2} | ∅ | B, AB |
| A | ∅ | ∅ | A |

* Tập phổ biến: {A, B, D, C, E, F, G, H, J, AB, AD, AG, BD, BH, BE, BJ, CD, CF, DF, DH, DJ, HJ, BDH, BDJ, BHJ, CDF, DHJ, BDHJ}.
* So sánh kết quả: Kết quả tìm được của 2 thuật toán là như nhau.
* Liệt kê tập phổ biến tối đại (**Maximal frequent itemsets**): AB, AD, AG, BE, CDF, BDHJ.
* Liệt kê các tập phổ biến đóng (**Closed frequent Itemsets**):

+ A: c(A) = i(t(A)) = i(200, 300, 400) = A.

+ B: c(B) = i(t(B)) = i(100, 300, 400) = B.

+ D: c(D) = i(t(D)) = i(100, 200, 300) = D.

+ AB: c(AB) = i(t(AB)) = i(300, 400) = AB.

+ AD: c(AD) = i(t(AD)) = i(200, 300) = AD.

+ AG: c(AG) = i(t(AG)) = i(200, 400) = AG.

+ BE: c(BE) = i(t(BE)) = i(100, 400) = BE.

+ CDF: c(CDF) = i(t(CDF)) = i(100, 200) = CDF.

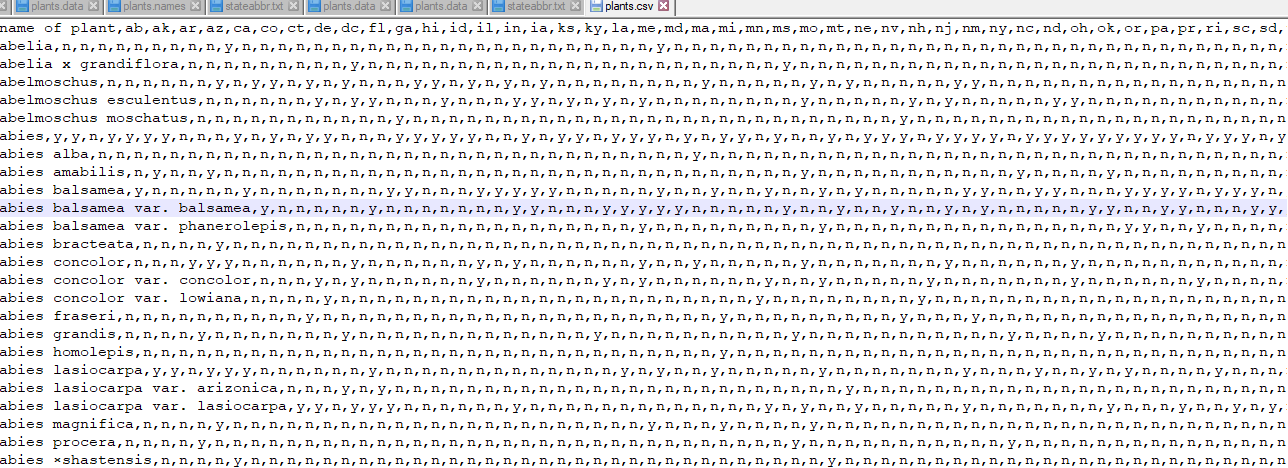
+ BDHJ: c(BDHJ) = i(t(BDHJ)) = i(100, 300) = BDHJ.

* Các tập phổ biến đóng: A, B, D, AB, AD, AG, BE, CDF, BDHJ.

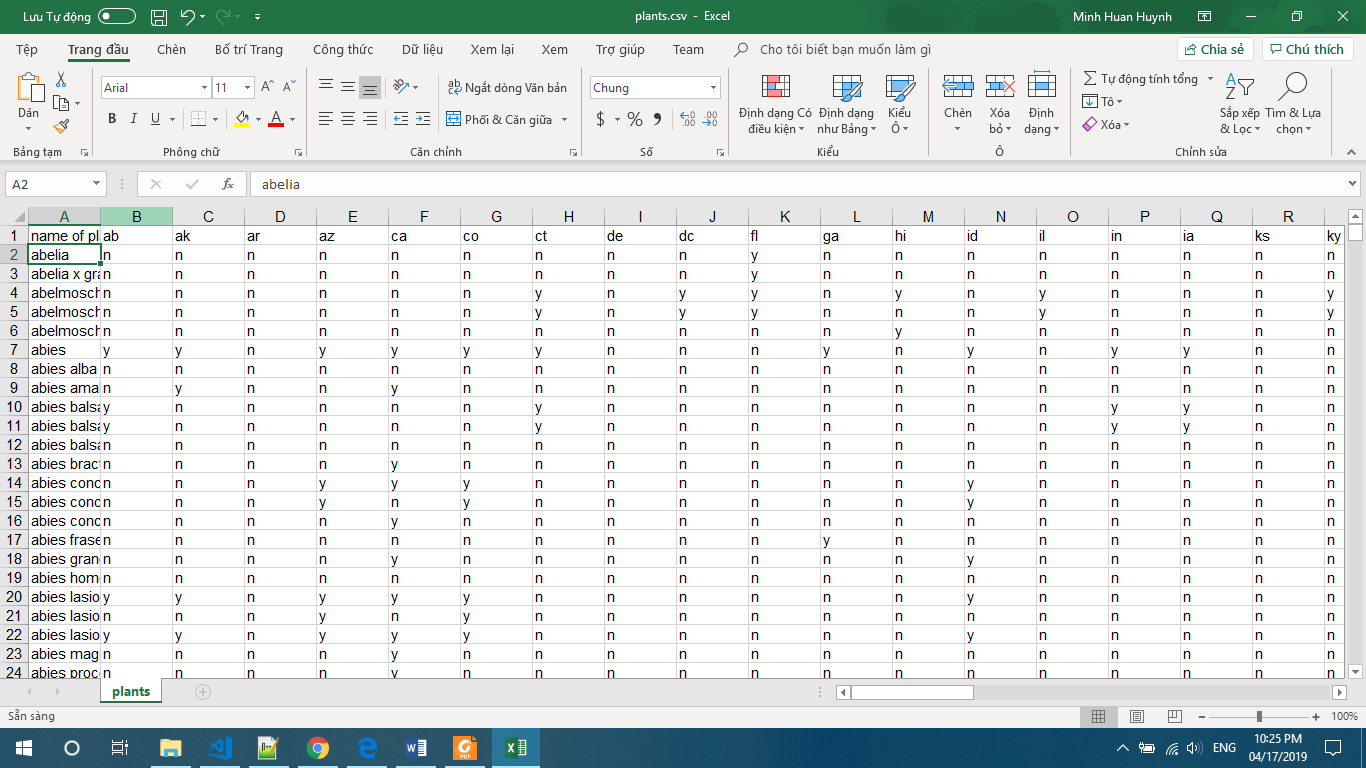
1. Tìm tất cả các luật kết hợp có dạng (item1 ^ item2 -> item3) thỏa mãn minsupp và minconf: BDH, BDJ, BHJ, CDF, DHJ, BDHJ

* BD -> H: conf(BD -> H) = 2/2 = 1.
* BH -> D: conf(BH -> D) = 2/2 = 1.
* DH -> B: conf(DH -> B) = 2/2 = 1.
* BD -> J: conf(BD -> J) = 2/2 = 1.
* BJ -> D: conf(BJ -> D) = 2/2 = 1.
* DJ -> B: conf(DJ -> B) = 2/2 = 1.
* BH -> J: conf(BH -> J) = 2/2 = 1.
* BJ -> H: conf(BJ -> H) = 2/2 = 1.
* HJ -> B: conf(HJ -> B) = 2/2 = 1.
* CD -> F: conf(CD -> F) = 2/2 = 1.
* CF -> D: conf(CF -> D) = 2/2 = 1.
* DF -> C: conf(DF -> C) = 2/2 = 1.
* DH -> J: conf(DH -> J) = 2/2 = 1.
* DJ -> H: conf(DJ -> H) = 2/2 = 1.
* HJ -> D: conf(HJ -> D) = 2/2 = 1.

1. Ứng dụng cải tiến của câu 1 vào việc tìm các luật kết hợp ở câu b thỏa mãn ngưỡng minconf. (So sánh hiệu quả về thời gian thực hiện kết quả ở câu b).
2. Thực hành:
3. Chuyển dữ liệu trong tập tin plants.data từ dạng giao dịch sang dạng nhị phân:

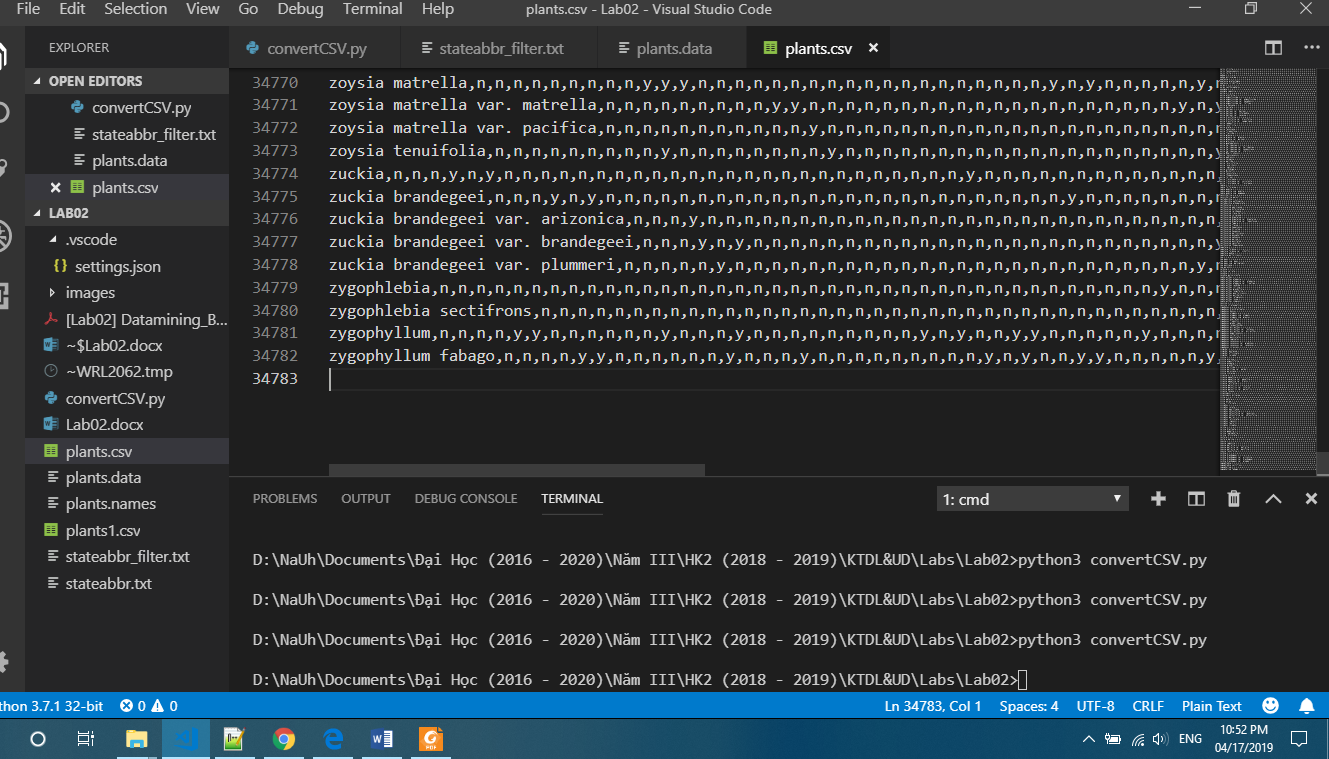


(Hình B1.1 - ảnh file plants.csv)



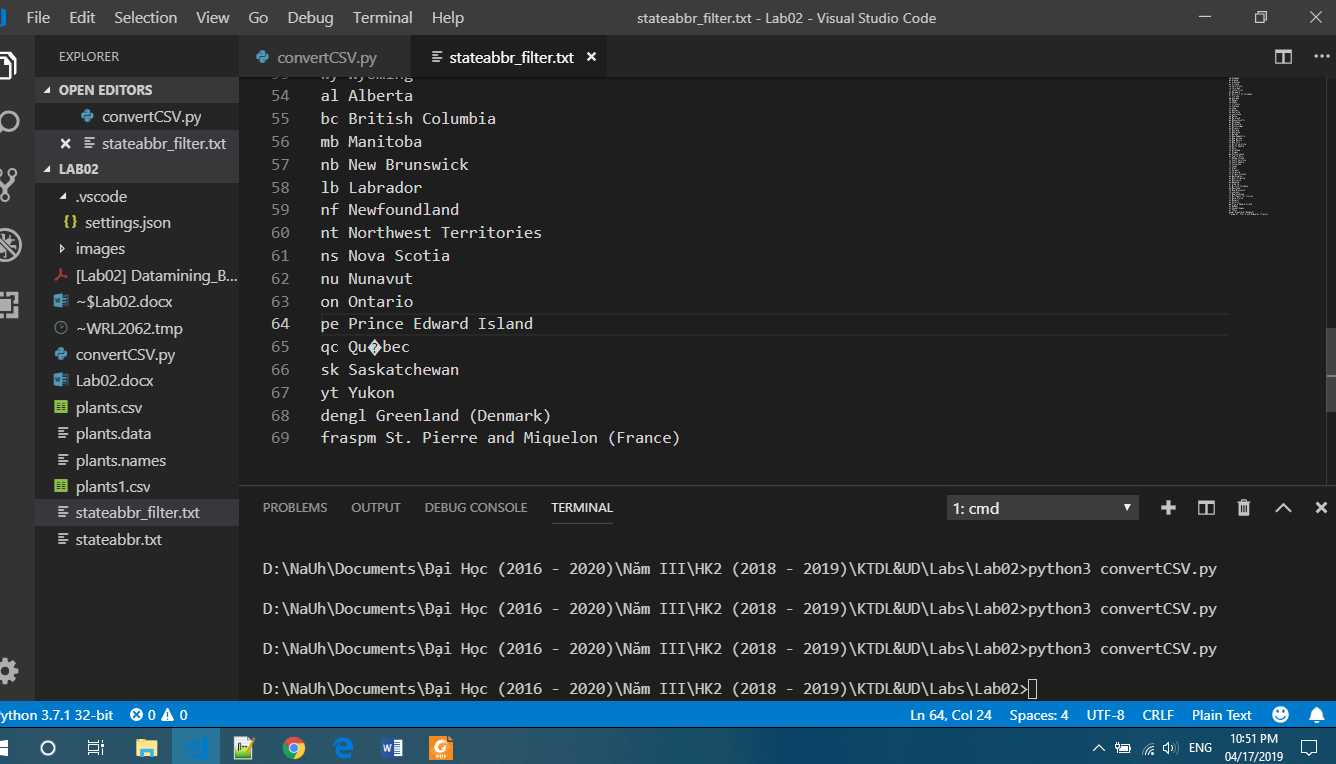
(Hình B1.2 - ảnh file plants.csv mở bằng excel)

1. Trả lời các câu hỏi:
2. Có tất cả 34781 loại cây.

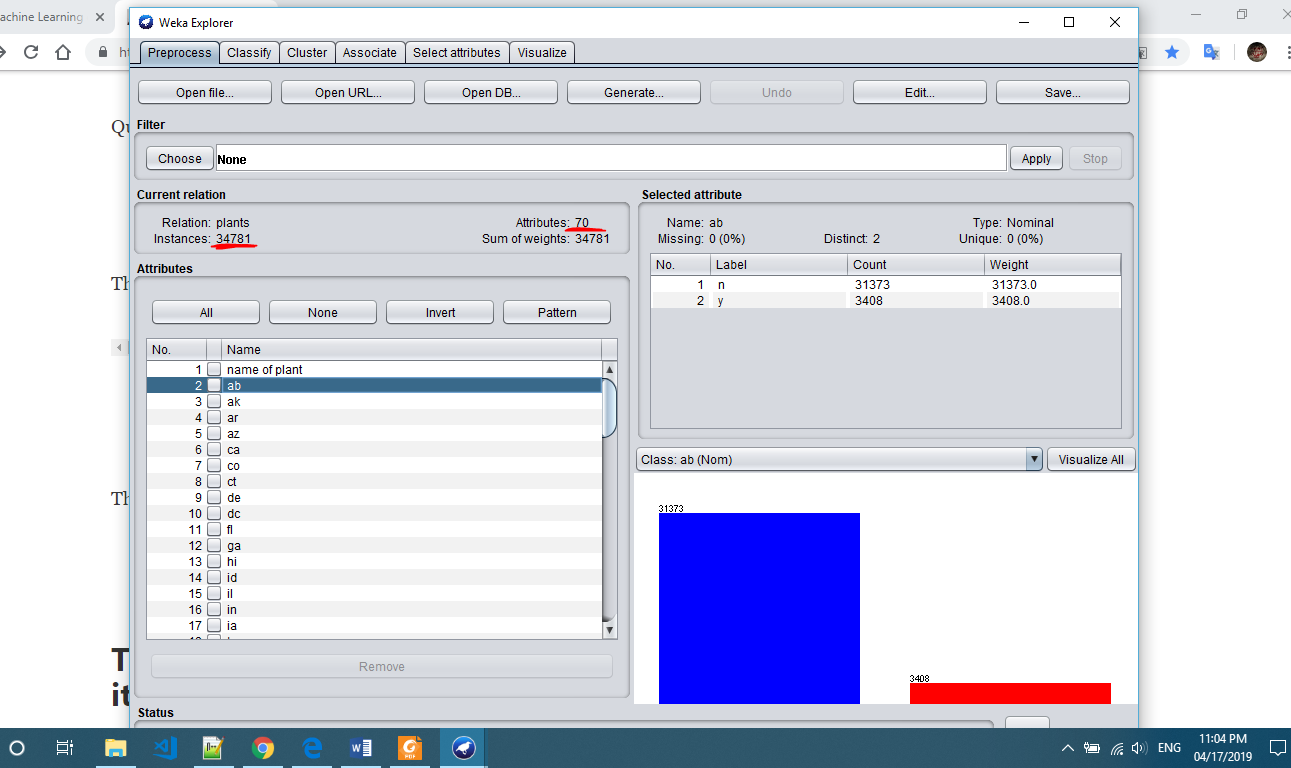


(Hình B2a)

1. Có tất cả 69 vùng phân bố.



(Hình B2b.1)

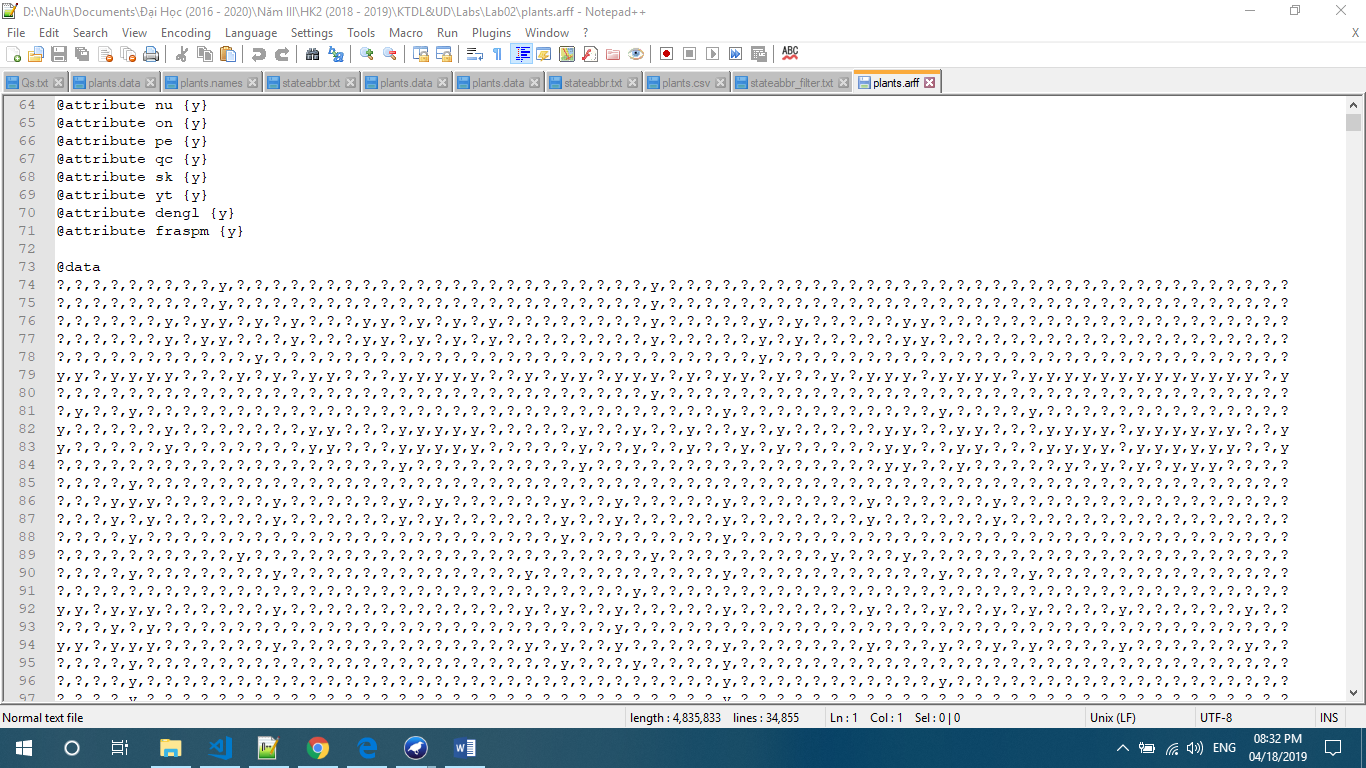


(Hình B2b.2)

1. Số loài cây trên mỗi vùng phân bố:

* ab Alabama: 3408.
* ak Alaska: 2969.
* ar Arkansas: 4610.
* az Arizona: 6778.
* ca California: 11676.
* co Colorado: 5465.
* ct Connecticut: 4391.
* de Delaware: 3630.
* dc District of Columbia: 3080.
* fl Florida: 6621.
* ga Georgia: 5942.
* hi Hawaii: 3804.
* id Idaho: 5129.
* il Illinois: 5167.
* in Indiana: 4440.
* ia Iowa: 3652.
* ks Kansas: 3869.
* ky Kentucky: 4555.
* la Louisiana: 5154.
* me Maine: 3969.
* md Maryland: 5108.
* ma Massachusetts: 4963.
* mi Michigan: 4734.
* mn Minnesota: 3929.
* ms Mississippi: 4815.
* mo Missouri: 4638.
* mt Montana: 4800.
* ne Nebraska: 3281.
* nv Nevada: 5670.
* nh New Hampshire: 3635.
* nj New Jersey: 4822.
* nm New Mexico: 6403.
* ny New York: 5773.
* nc North Carolina: 5926.
* nd North Dakota: 2682.
* oh Ohio: 4772.
* ok Oklahoma: 4651.
* or Oregon: 7028.
* pa Pennsylvania: 5474.
* pr Puerto Rico: 4781.
* ri Rhode Island: 3295.
* sc South Carolina: 5432.
* sd South Dakota: 3185.
* tn Tennessee: 4900.
* tx Texas: 8483.
* ut Utah: 6041.
* vt Vermont: 3713.
* va Virginia: 5638.
* vi Virgin Islands: 2185.
* wa Washington: 5654.
* wv West Virginia: 4062.
* wi Wisconsin: 4321.
* wy Wyoming: 4710.
* al Alberta: 5702.
* bc British Columbia: 4875.
* mb Manitoba: 3023.
* nb New Brunswick: 2856.
* lb Labrador: 1433.
* nf Newfoundland: 2188.
* nt Northwest Territories: 2024.
* ns Nova Scotia: 2844.
* nu Nunavut: 979.
* on Ontario: 5068.
* pe Prince Edward Island: 1841.
* qc Qu�bec: 4272.
* sk Saskatchewan: 2846.
* yt Yukon: 2100.
* dengl Greenland (Denmark): 479.
* fraspm St. Pierre and Miquelon (France): 1210.

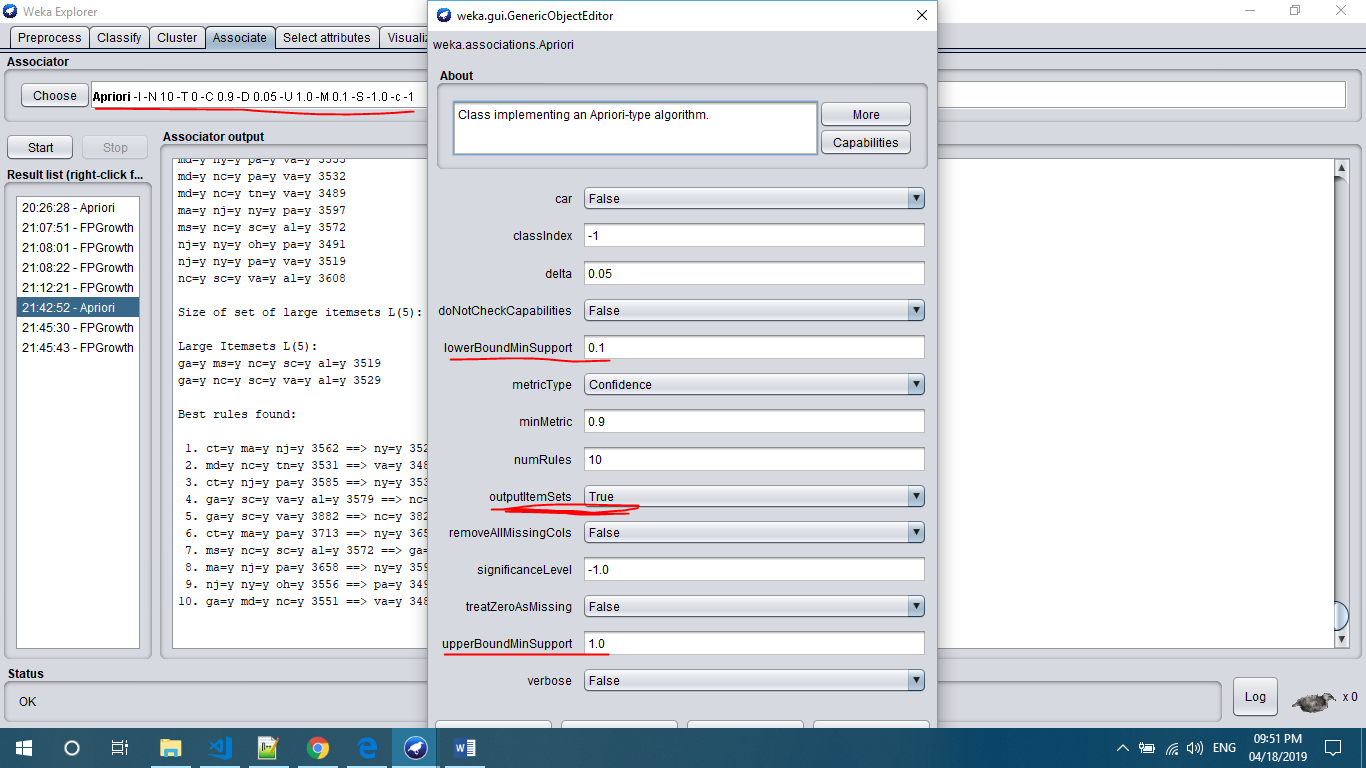
1. Vùng phân bố ít loài cây nhất dengl - Greenland (Denmark): 479 loài với tỉ lệ 1,377%.
2. Vùng phân bố nhiều loài cây nhất ca – California: 11676 loài với tỉ lệ 33.57%.
3. Trung bình một vùng phân bố có 4313 loài cây.
4. Áp dụng thuật giải Apriori trên dữ liệu này. Xử lý file sang định dạng plants.arff:



(Hình B3)

1. Khai thác tập phổ biến:

* Yêu cầu: sử dụng thuật toán Apriori trong Weka để khai thác tập hạng mục có độ phổ biến từ 0.1 trở lên
* Các bước thực hiện: chọn option

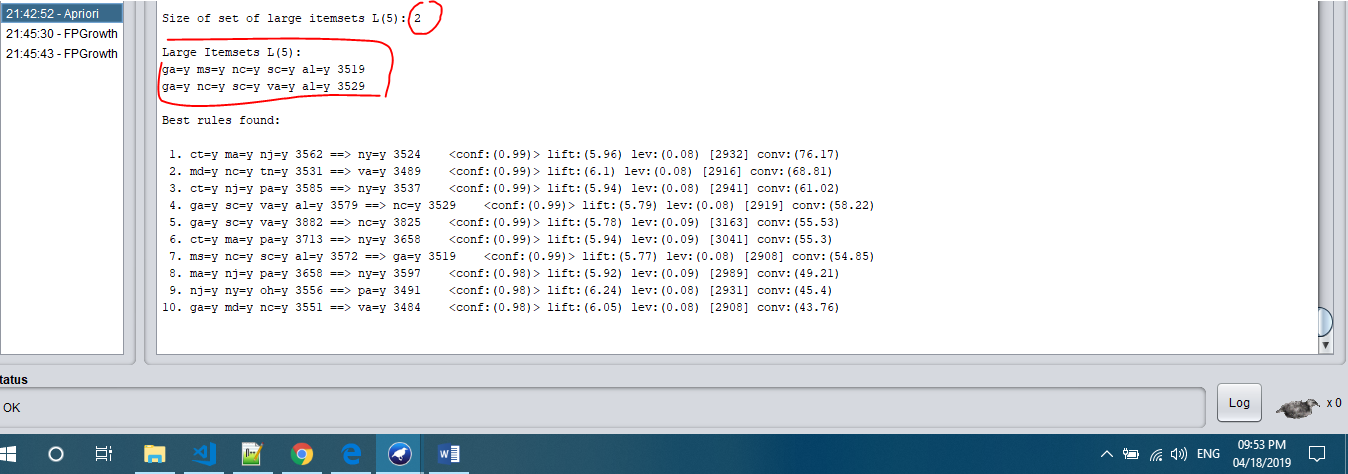


(Hình B4)

|  |  |
| --- | --- |
| 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. Khai thác luật kết hợp:

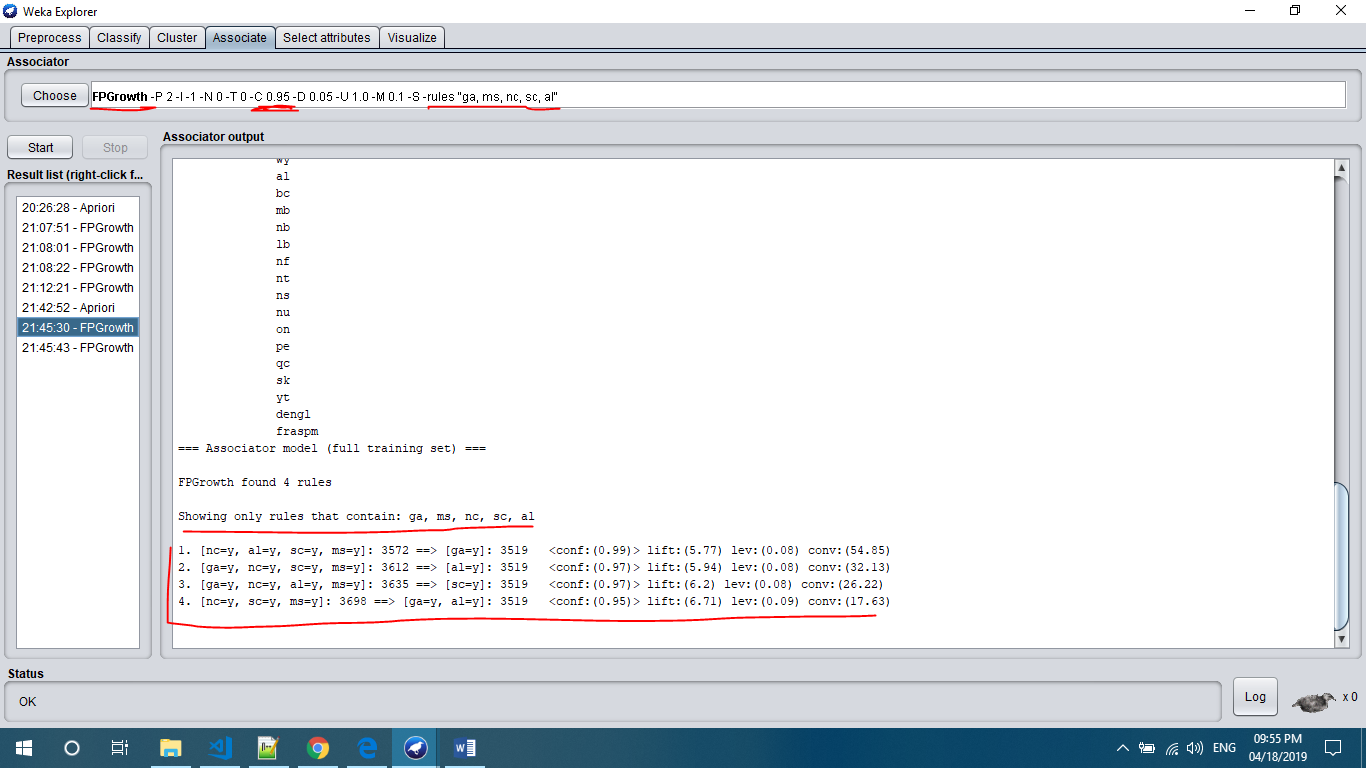
* Với mỗi tập phổ biến L5 theo kết quả của câu 4:



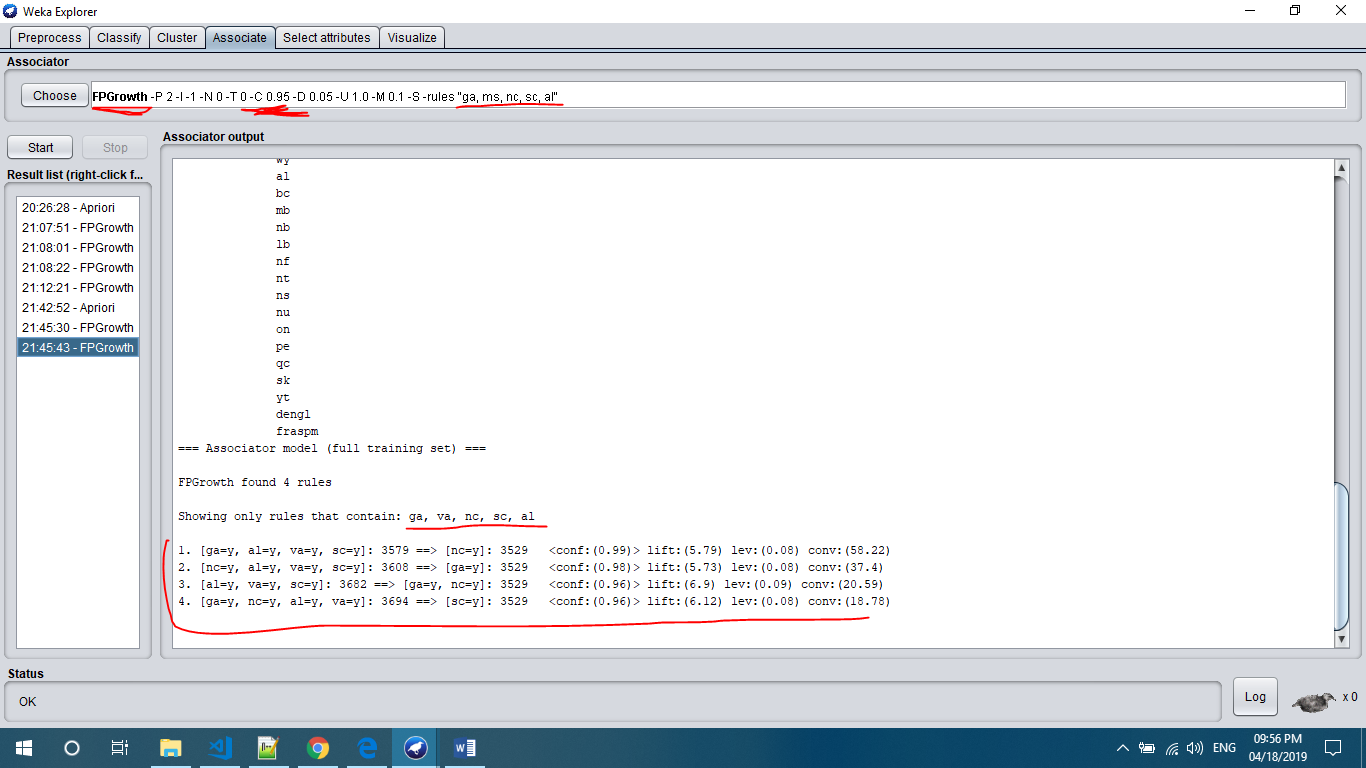
(Hình B5.1)

* Kết quả:

|  |  |
| --- | --- |
| Tập các hạng mục phổ biến | Số lượng luật |
| ga =y, al=y, va=y, sc=y, nc=y | 4 |
| ga=y, al=y, ms=y, sc=y, nc=y | 4 |



(Hình B5.2 các luật rút từ tập {ga, ms, nc, sc, al})



(Hình B5.3 Các luật rút ra từ tập {ga, va, nc, sc, al})