TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO
BÀI TẬP THỰC HÀNH TUẦN 1
Khai thác dữ liệu Web

Giảng viên hướng dẫn : Khoa Phó Ngọc Đăng

Sinh viên thực hiện : Vũ Mạnh Hùng

Mã số sinh viên : 1461390

Lớp : 15CK3

Ca : 1 – C6

TP HCM, tháng 3 năm 2017

Bài tập 1) Apriori & FP-Growth:

a) Minsup = 35% => 0.35

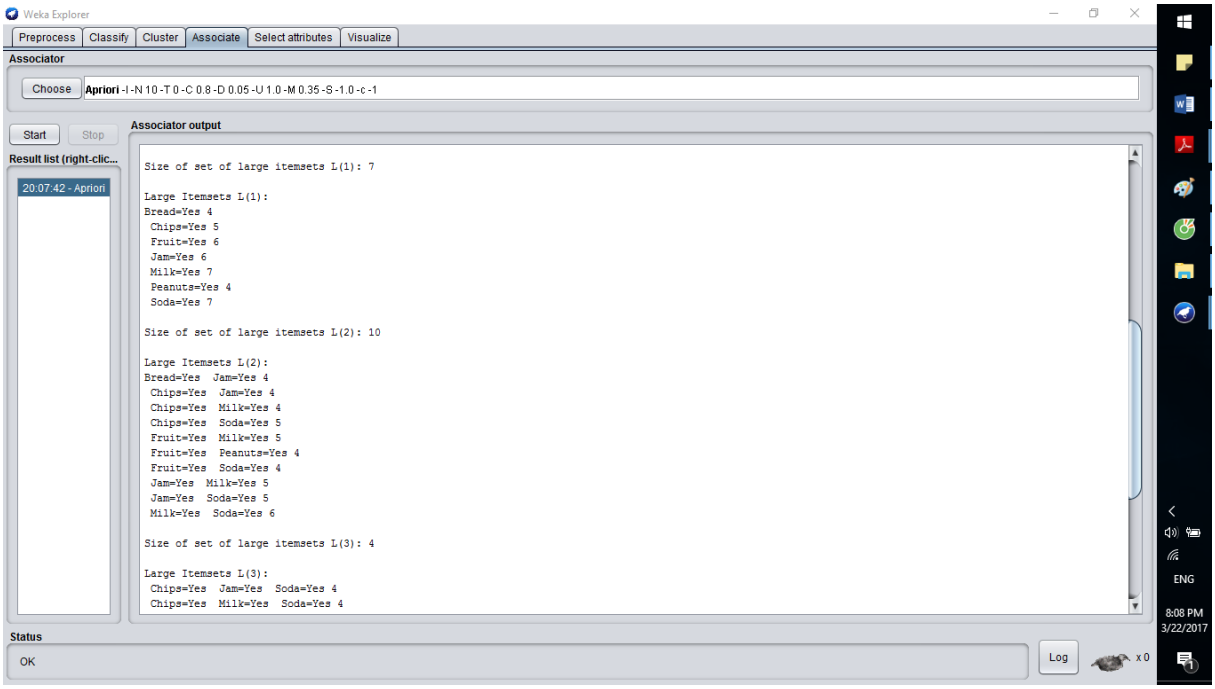
=> Minsupcount = 0.35 × 10 = 3.5 = 4

F_1	
Item	Sup
Bread	4
Chips	5
Fruit	6
Jam	6
Milk	7
Peanuts	4
Soda	7

F_2	
Item	Sup
Bread, Jam	4
Chips, Jam	4
Chips, Milk	4
Chips, Soda	5
Fruit, Milk	5
Fruit, Peanuts	4
Fruit, Soda	4
Jam, Milk	5
Jam, Soda	5
Milk, Soda	6

F_3	
Item	Sup
Chips, Jam, Soda	4
Chips, Milk, Soda	4
Fruit, Milk, Soda	4

Jam, Milk, Soda	4



b) Liệt kê tập phổ biến tối đại, tập phổ biến đóng:

Tập phổ biến đóng	Tập phổ biến tối đại
Fruit	Bead, Jam
Jam	Chips, Soda
Milk	Fruit, Peanuts
Soda	Fruit, Soda
Bead, Jam	Jam, Soda
Fruit, Milk	Milk, Soda
Chips, Soda	
Fruit, Peanuts	
Fruit, Soda	
Jam, Soda	
Milk, Soda	
Chips, Jam, Soda	
Chips, Milk, Soda	
Fruit, Milk, Soda	
Jam, Milk, Soda	

c) Liệt kê tất cả các luật kết hợp có dạng {item1, item2} → item3 thỏa ngưỡng minsup và minconf đã cho:

❖ Chips, Jam, Soda

- Chips, Jam → Soda 4/4
- Chips, Soda →Jam 4/5
- Jam, Soda →Chips 4/5

❖ Chips, Milk, Soda

- Chips, Soda → Milk 4/5
- Chips, Milk → Soda 4/4

❖ Fruit, Milk, Soda

- Fruit, Milk → Soda 4/5
- Fruit, Soda → Milk 4/4

❖ Jam, Milk, Soda

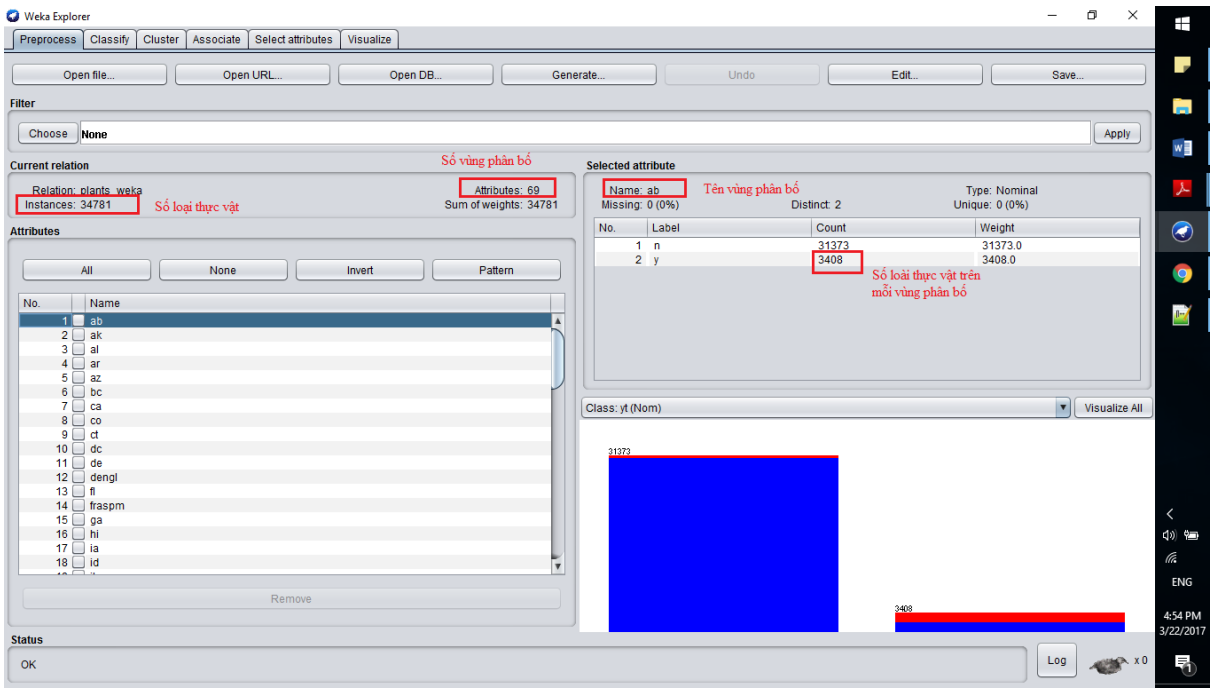
- Jam, Milk → Soda 4/5
- Jam, Soda → Milk 4/5

Bài tập 2) Thực hành với công cụ Weka:

a) Hãy chuyển dữ liệu trong tập tin plants.data từ dạng giao dịch sang dạng nhị phân như sau:

Bài làm trong file plants.csv

b) Sử dụng công cụ WEKA và trả lời câu hỏi sau:



- Có tất cả bao nhiêu loài thực vật. 34781
- Có tất cả bao nhiêu vùng phân bố: 69
- Số loài thực vật trên mỗi vùng phân bố.

STT	Vùng phân bố	Số loài thực vật
1	ab	3408
2	ak	2969
3	al	5702
4	ar	4610
5	az	6778
6	bc	4875
7	ca	11676
8	co	5465
9	ct	4391
10	dc	3080
11	de	3630
12	dengl	479
13	fl	28160

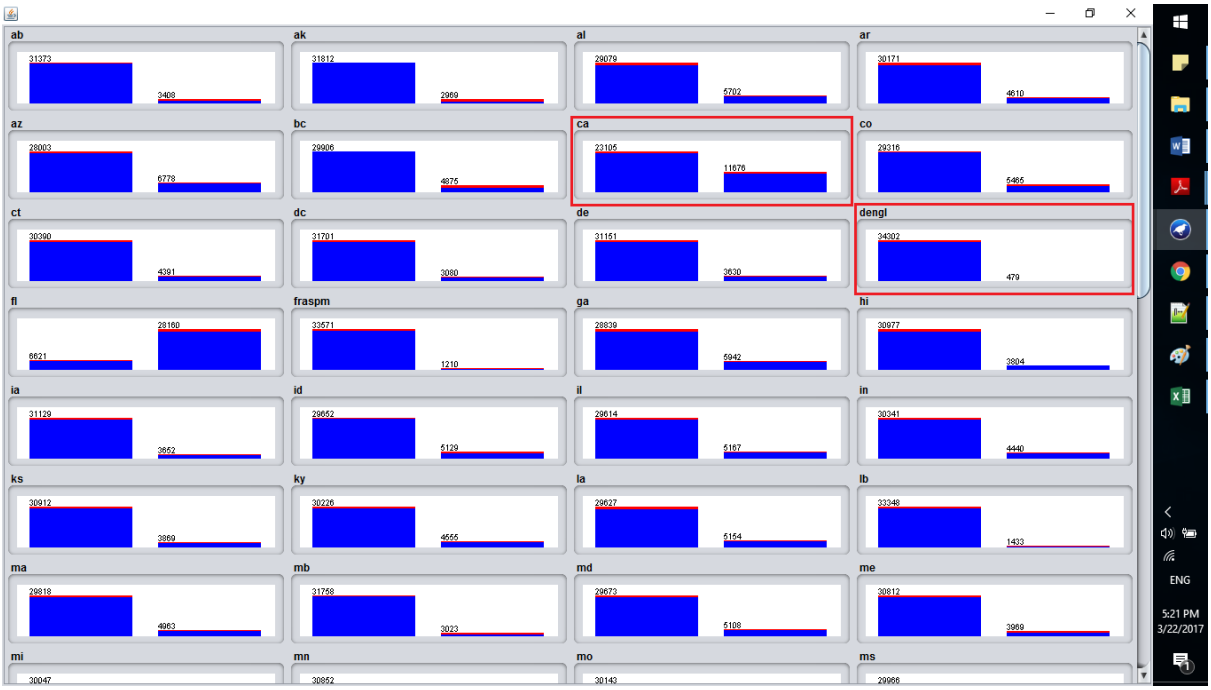
14	fraspm	1210
15	ga	5942
16	hi	3804
17	ia	3652
18	id	5129
19	il	5167
20	in	4440
21	ks	3869
22	ky	4555
23	la	5145
24	lb	1433
25	ma	4963
26	mb	3023
27	md	5108
28	me	3969
29	mi	4734
30	mn	3929

31	mo	4638
32	ms	4815
33	mt	4800
34	nb	2856
35	nc	28855
36	nd	2682
37	ne	3281
38	nf	2188
39	nh	3635
40	nj	4822
41	nm	6403
42	ns	2844
43	nt	2024
44	nu	979
45	nv	5670
46	ny	5773
47	oh	4772

48	ok	4651
49	on	5068
50	or	7028
51	pa	5474
52	pr	4781
53	Pr	4781
54	qc	4272
55	ri	3295
56	sc	5432
57	sd	3185
58	sk	2846
59	tn	4900
60	tx	8483
61	ut	6041
62	va	5638
63	vi	2185
64	vt	3713

65	wa	5654
66	wi	4321
67	wv	4062
68	wy	4710
69	yt	2100

- Vùng phân bố có ít loài cây nhất, cho biết số lượng tương ứng: dengl - 479
- Vùng phân bố có nhiều loài cây nhất, cho biết số lượng tương ứng. ca – 11676



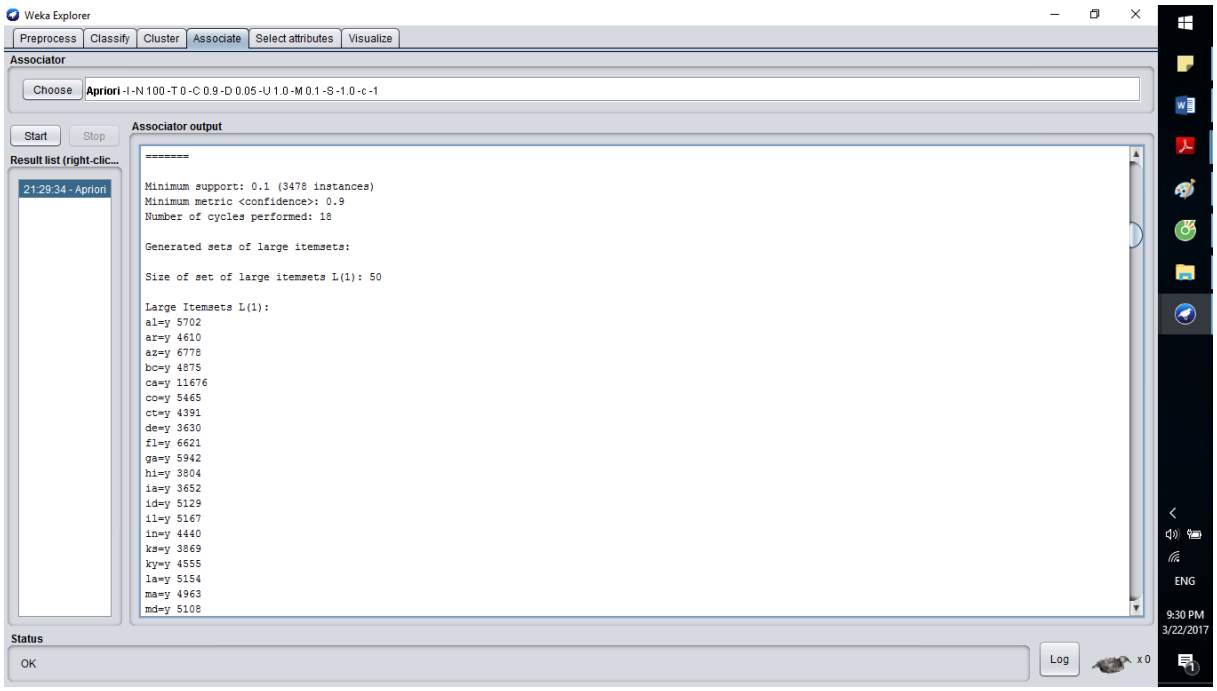
- Trung bình một vùng phân bố có bao nhiêu loài cây: 4412

c) Điều chỉnh nội dung của tập tin để khai thác mẫu phổ biến như sau:

Bài làm trong file plants.arff

d) Khai thác tập phổ biến:

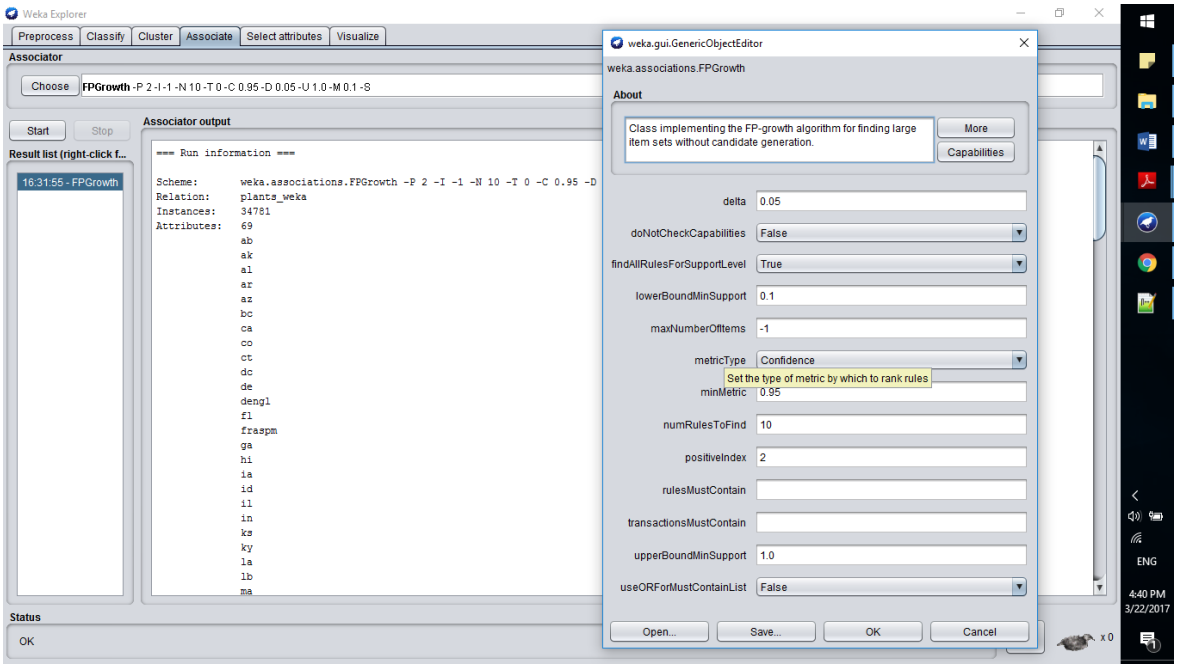
Kích thước	Số lượng tập phổ biến
1 hạng mục	50
2 hạng mục	168
3 hạng mục	116
4 hạng mục	25
5 hạng mục	2



File kết quả lưu trong F1.doc

e) Khai thác luật kết hợp:

Tập hạng mục phổ biến	Số lượng luật



File kết quả lưu trong AR.doc