Xiaobo Zhang
4/28/2015

Hw5

Part 1

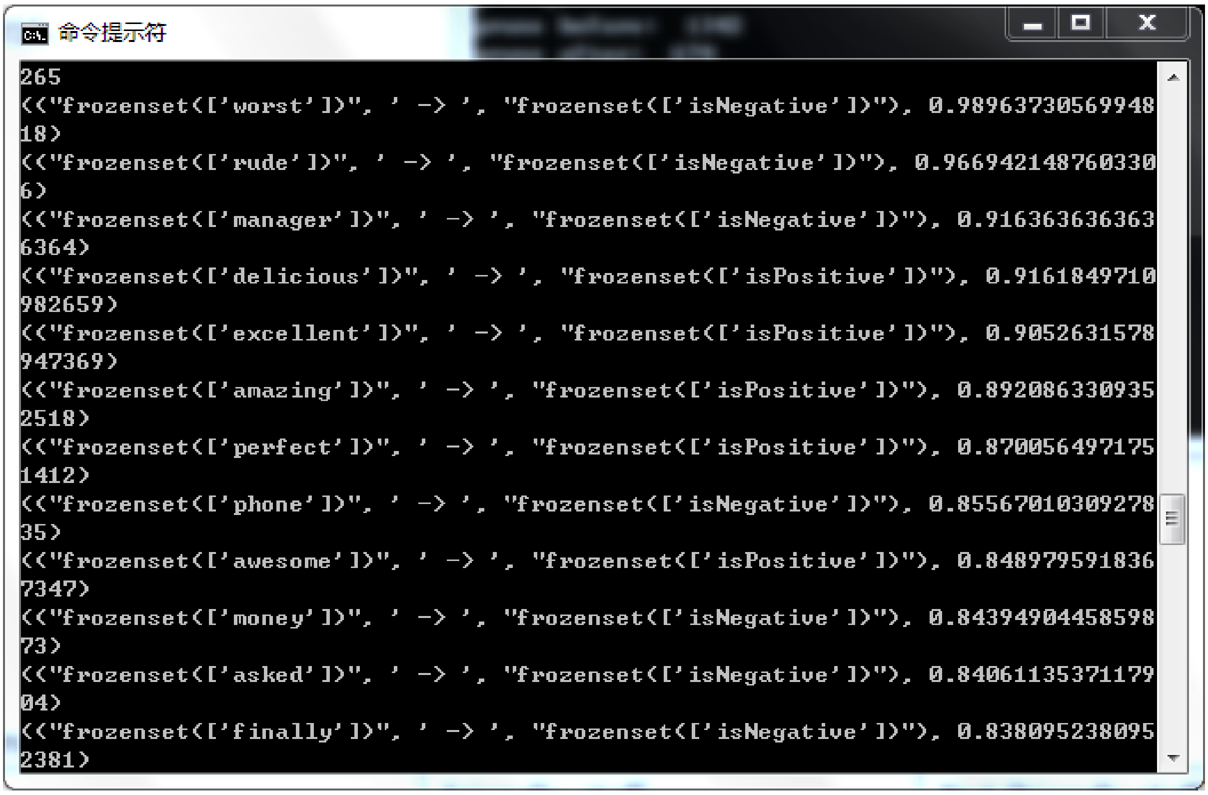
a) The pattern space should be: $\binom{2002}{1}+\binom{2002}{2}+\binom{2002}{3}$ = 2002 + 2003001 + 1335334000 = 1337339003

b) The frequent item is 609, and the infrequent will be 1337339003 - 609

c) The prune ratio is (1337339003 - 609) / 1337339003 = 0.99

d) False alarm rate: (609 - 205)/609 = 0.66

e) Top 30 rules:




```
265
(("frozenset(['worst'])", ' -> ', "frozenset(['isNegative'])"), 0.98963730569948
18)
(("frozenset(['rude'])", ' -> ', "frozenset(['isNegative'])"), 0.966942148760330
6)
(("frozenset(['manager'])", ' -> ', "frozenset(['isNegative'])"), 0.916363636363
6364)
(("frozenset(['delicious'])", ' -> ', "frozenset(['isPositive'])"), 0.9161849710
982659)
(("frozenset(['excellent'])", ' -> ', "frozenset(['isPositive'])"), 0.9052631578
947369)
(("frozenset(['amazing'])", ' -> ', "frozenset(['isPositive'])"), 0.892086330935
2518)
(("frozenset(['perfect'])", ' -> ', "frozenset(['isPositive'])"), 0.870056497175
1412)
(("frozenset(['phone'])", ' -> ', "frozenset(['isNegative'])"), 0.85567010309278
35)
(("frozenset(['awesome'])", ' -> ', "frozenset(['isPositive'])"), 0.848979591836
7347)
(("frozenset(['money'])", ' -> ', "frozenset(['isNegative'])"), 0.84394904458598
73)
(("frozenset(['asked'])", ' -> ', "frozenset(['isNegative'])"), 0.84061135371179
04)
(("frozenset(['finally'])", ' -> ', "frozenset(['isNegative'])"), 0.838095238095
2381)
```

命令提示符

```
2381)
(("frozenset(['friendly', 'staff'])", ' -> ', "frozenset(['isPositive'])"), 0.83
68421052631579)
(("frozenset(['later'])", ' -> ', "frozenset(['isNegative'])"), 0.83571428571428
57)
(("frozenset(['favorite'])", ' -> ', "frozenset(['isPositive'])"), 0.82993197278
91157)
(("frozenset(['friendly'])", ' -> ', "frozenset(['isPositive'])"), 0.82320441988
95028)
(("frozenset(['minutes'])", ' -> ', "frozenset(['isNegative'])"), 0.814531548757
1702)
(("frozenset(['customers'])", ' -> ', "frozenset(['isNegative'])"), 0.8088888888
888889)
(("frozenset(['15'])", ' -> ', "frozenset(['isNegative'])"), 0.7980295566502463)

(("frozenset(['customer'])", ' -> ', "frozenset(['isNegative'])"), 0.79389312977
09924)
(("frozenset(['call'])", ' -> ', "frozenset(['isNegative'])"), 0.789655172413793
1)
(("frozenset(['should'])", ' -> ', "frozenset(['isNegative'])"), 0.7884615384615
384)
(("frozenset(['nothing'])", ' -> ', "frozenset(['isNegative'])"), 0.788461538461
5384)
(("frozenset(['waiting'])", ' -> ', "frozenset(['isNegative'])"), 0.786764705882
3529)
```

命令提示符

```
(("frozenset(['customer'])", ' -> ', "frozenset(['isNegative'])"), 0.79389312977
09924)
(("frozenset(['call'])", ' -> ', "frozenset(['isNegative'])"), 0.789655172413793
1)
(("frozenset(['should'])", ' -> ', "frozenset(['isNegative'])"), 0.7884615384615
384)
(("frozenset(['nothing'])", ' -> ', "frozenset(['isNegative'])"), 0.788461538461
5384)
(("frozenset(['waiting'])", ' -> ', "frozenset(['isNegative'])"), 0.786764705882
3529)
(("frozenset(['left'])", ' -> ', "frozenset(['isNegative'])"), 0.782608695652174
)
(("frozenset(['called'])", ' -> ', "frozenset(['isNegative'])"), 0.7808641975308
642)
(("frozenset(['point'])", ' -> ', "frozenset(['isNegative'])"), 0.77889447236180
91)
(("frozenset(['love'])", ' -> ', "frozenset(['isPositive'])"), 0.77358490566037
8)
(("frozenset(['20'])", ' -> ', "frozenset(['isNegative'])"), 0.7766990291262136)

(("frozenset(['him'])", ' -> ', "frozenset(['isNegative'])"), 0.7762237762237763
)

C:\Users\Xiaobo Zhang\Desktop\xiaobo_zhang>
```

According these 30 rules, it sounds reasonable. For example, friendly, staff to positive, rude to negative, worst to negative, and so on. These rules can reflect the association between two sides very well.

Part 2

Chi-square formula: $x^2 = \sum_i^n \frac{(expected_i - observered_i)^2}{expected_i}$

Rules friendly, staff-> isPositive

	Friendly ,staff	Not (friendly,staff)
isPositive	159	2341
Not isPositive	31	2469

Score: 89.63

Accuracy: 0.5256

Generalize: Rules friendly-> isPositive

	Friendly	Not Friendly
Positive	447	2053
Not positive	96	2404

Score: 254.53 Accuracy: 0.5702

As the rule become more general, there are more satisfied item sets in the contingency table, the accuracy will be larger than before.

Specialize: Rules friendly, staff, favorite -> isPositive

	friendly, staff, favorite	Not friendly, staff, favorite
isPositive	15	2485
Not isPositive	0	2500

Score: 15.04 Accuracy: 0.5053

As the rule becomes more specialize, there are less satisfied item sets in the contingency table than previous item size.

When the rule more specialize, the accuracy will be lower and lower, the accuracy can be alternative for threshold evaluation to reduce the space size. It is similar to Aprior Algorithm that the more attribute in the rules, it is more impossible to generate rules. If one subsequent rules is small, it can be also determine that other 2 subsequent or more rules will be much smaller.

Part3

1. Top 30 rules:

命令提示符

```
656
support count after:  1
30
(("frozenset(['isPositive', 'staff'])", ' -> ', "frozenset(['friendly'])"), 413.0395867343834)
(("frozenset(['staff'])", ' -> ', "frozenset(['friendly'])"), 262.09789455077504)
(("frozenset(['friendly'])", ' -> ', "frozenset(['staff'])"), 262.09789455077504)
(("frozenset(['delicious'])", ' -> ', "frozenset(['isPositive'])"), 257.5446318786003)
(("frozenset(['isPositive'])", ' -> ', "frozenset(['delicious'])"), 257.5446318786003)
(("frozenset(['isPositive'])", ' -> ', "frozenset(['friendly'])"), 254.53163872832727)
(("frozenset(['friendly'])", ' -> ', "frozenset(['isPositive'])"), 254.53163872832727)
(("frozenset(['better'])", ' -> ', "frozenset(['than'])"), 236.54319334674955)
(("frozenset(['than'])", ' -> ', "frozenset(['better'])"), 236.54319334674955)
(("frozenset(['asked'])", ' -> ', "frozenset(['isNegative'])"), 233.9734530120621)
(("frozenset(['isNegative'])", ' -> ', "frozenset(['asked'])"), 233.9734530120621)
(("frozenset(['isNegative'])", ' -> ', "frozenset(['minutes'])"), 231.1388866229819)
```

命令提示符

```
(("frozenset(['isNegative'])", ' -> ', "frozenset(['minutes'])"), 231.1388866229819)
(("frozenset(['minutes'])", ' -> ', "frozenset(['isNegative'])"), 231.1388866229819)
(("frozenset(['isPositive', 'friendly'])", ' -> ', "frozenset(['staff'])"), 222.4719008965333)
(("frozenset(['isNegative'])", ' -> ', "frozenset(['rude'])"), 221.79261374492373)
(("frozenset(['rude'])", ' -> ', "frozenset(['isNegative'])"), 221.79261374492373)
(("frozenset(['came'])", ' -> ', "frozenset(['ordered'])"), 205.48268752387426)
(("frozenset(['ordered'])", ' -> ', "frozenset(['came'])"), 205.48268752387423)
(("frozenset(['isNegative'])", ' -> ', "frozenset(['manager'])"), 201.7931697931698)
(("frozenset(['manager'])", ' -> ', "frozenset(['isNegative'])"), 201.7931697931698)
(("frozenset(['worst'])", ' -> ', "frozenset(['isNegative'])"), 192.5139396238861 6)
(("frozenset(['isNegative'])", ' -> ', "frozenset(['worst'])"), 192.5139396238861 3)
(("frozenset(['isPositive'])", ' -> ', "frozenset(['again'])"), 184.36705616829073)
(("frozenset(['again'])", ' -> ', "frozenset(['isPositive'])"), 184.36705616829073)
(("frozenset(['again'])", ' -> ', "frozenset(['isNegative'])"), 184.367056168290
```

2. In association rules, the chi square score will take only one test, so it may be different with score take more than 100 times or more. So the final result will not reflect the truth of the data, which means the chi-square score will not at right position in normal distribution of tests. Therefore, the result from association rules will be incorrect.
3. In Bonferroni correction, decreasing the significant level smaller by alpha/comparison #, which can

exclude some unrelated data.

4. The original rules will create 445 rules, after using Bonferroni correction, the size of rules reduces to 337.