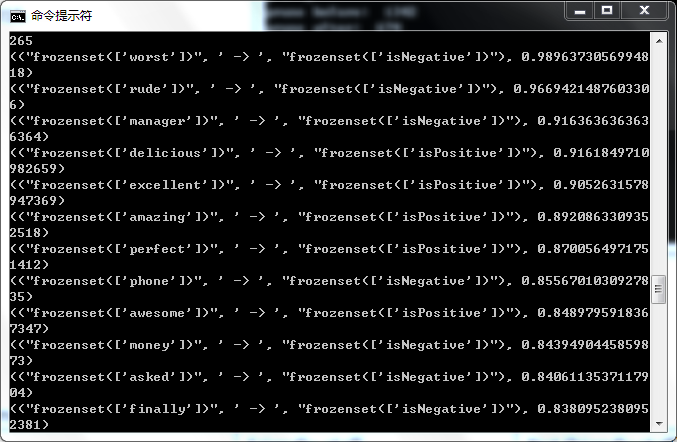
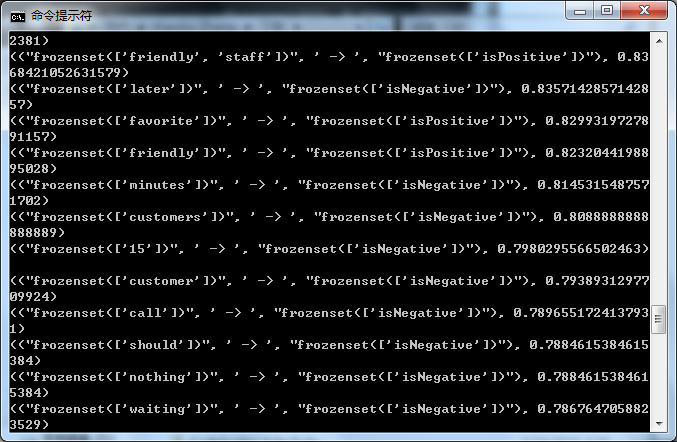
Xiaobo Zhang

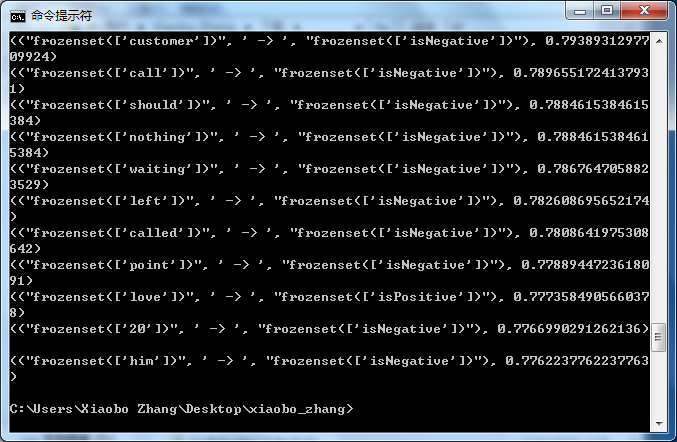
4/28/2015

Hw5

Part 1

1. The pattern space should be: = 2002 + 2003001 + 1335334000 = 1337339003
2. The frequent item is 609, and the infrequent will be 1337339003 - 609
3. The prune ratio is (1337339003 - 609) / 1337339003 = 0.99
4. False alarm rate: (609 - 205)/609 = 0.66
5. Top 30 rules:  
   





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:

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 |

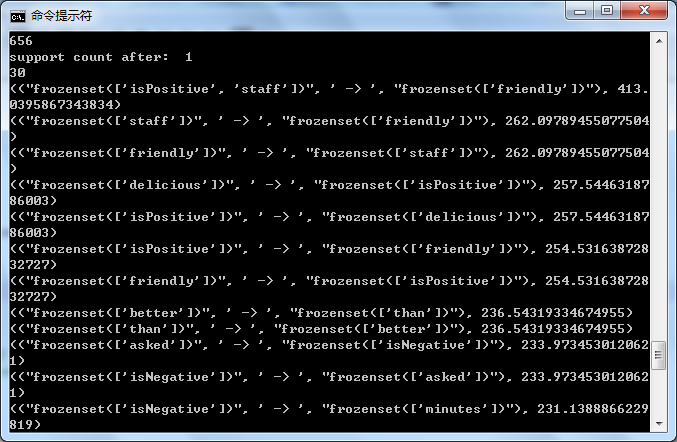
Score: 15.04 Accuracy: 0.5053

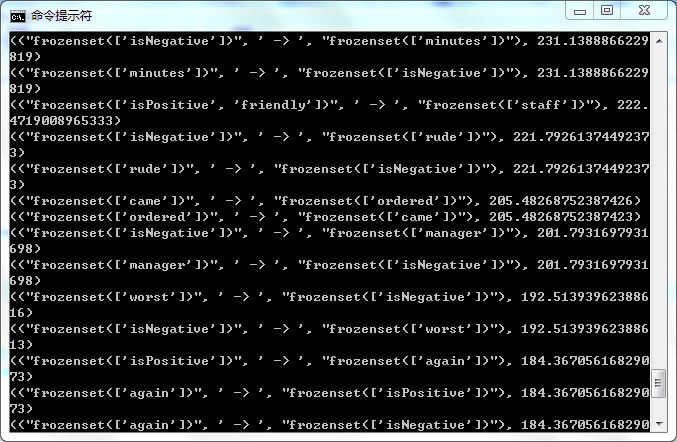
As the rule becomes more specialize, there are less satisfied item sets in the contingency table than previous item size.

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:





1. 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.
2. In Bonferroni correction, decreasing the significant level smaller by alpha/comparison #, which can exclude some unrelated data.
3. The original rules will create 445 rules, after using Bonferroni correction, the size of rules reduces to 337.