Project1: Association Analysis

E54036219 張富嘉

I. Software usage:

I have written the project using python, and have used the IBM Data Generator to generate association data. The usages are as follow:

```
./fpGrowth.py -f <datafile> -s minSupport -c minConfidence
```

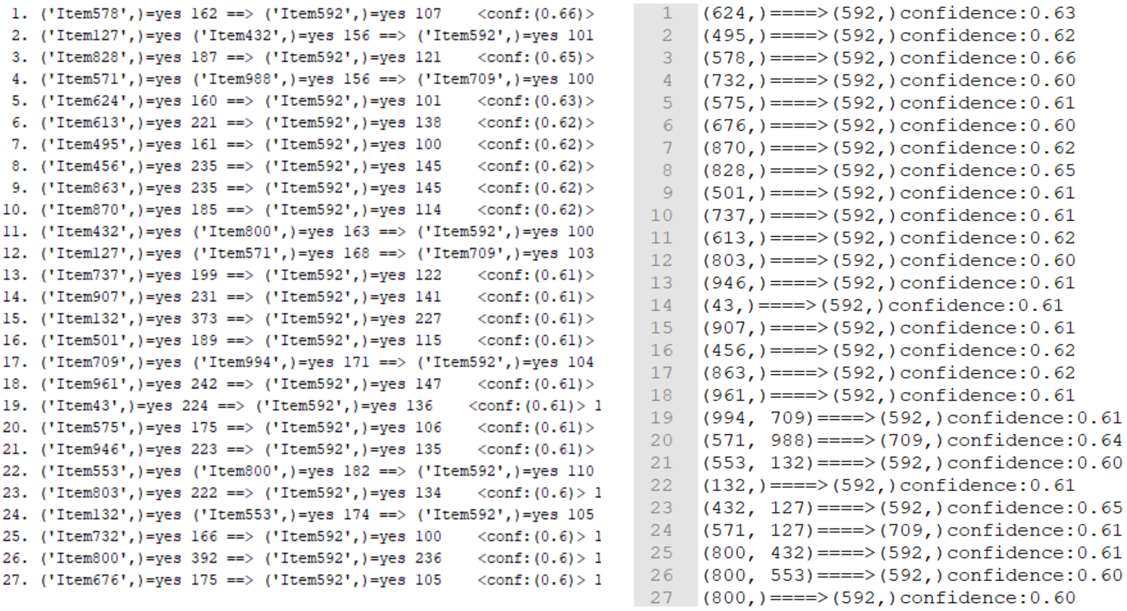
`<datafile>` should be in .data format that is generated by IBM Data Generator. `minSupport` is the number of minimum support, which is a positive integer. `minConfidence` is the number of minimum confidence, which is a floating number between 0 and 1.

II. Result and Output:

I have implemented a FP-Tree data structure in this project. The main purpose of constructing FP-Tree is to find out the frequent itemset more efficiently comparing to normal Apriori algorithm through mining on this tree structure recursively. After the frequent itemset is accomplished, we can easily generate the association rules according to the minimum confidence. Once the program is executed successfully, two files named in “freq.txt” and “rules.txt” are created under the folder, containing frequent-itemset and association rules respectively.

III. Compare to WEKA:

By comparing the result from my program and from WEKA(Apriori algo.), for the same testing data(data1), we can find that the frequent-itemset and association rules are exactly the same under minSupport=100, and minConfidence=0.6. However, the execution time varies greatly. WEKA spent couples of minutes, while FP-Growth spent only a few seconds. Such result also reflects previous assumption.



```
 1. ('Item578',)=yes 162 ==> ('Item592',)=yes 107    <conf:(0.66)>
 2. ('Item127',)=yes ('Item432',)=yes 156 ==> ('Item592',)=yes 101
 3. ('Item828',)=yes 187 ==> ('Item592',)=yes 121    <conf:(0.65)>
 4. ('Item571',)=yes ('Item988',)=yes 156 ==> ('Item709',)=yes 100
 5. ('Item624',)=yes 160 ==> ('Item592',)=yes 101    <conf:(0.63)>
 6. ('Item613',)=yes 221 ==> ('Item592',)=yes 138    <conf:(0.62)>
 7. ('Item495',)=yes 161 ==> ('Item592',)=yes 100    <conf:(0.62)>
 8. ('Item456',)=yes 235 ==> ('Item592',)=yes 145    <conf:(0.62)>
 9. ('Item863',)=yes 235 ==> ('Item592',)=yes 145    <conf:(0.62)>
10. ('Item870',)=yes 185 ==> ('Item592',)=yes 114    <conf:(0.62)>
11. ('Item432',)=yes ('Item800',)=yes 163 ==> ('Item592',)=yes 100
12. ('Item127',)=yes ('Item571',)=yes 168 ==> ('Item709',)=yes 103
13. ('Item737',)=yes 199 ==> ('Item592',)=yes 122    <conf:(0.61)>
14. ('Item907',)=yes 231 ==> ('Item592',)=yes 141    <conf:(0.61)>
15. ('Item132',)=yes 373 ==> ('Item592',)=yes 227    <conf:(0.61)>
16. ('Item501',)=yes 189 ==> ('Item592',)=yes 115    <conf:(0.61)>
17. ('Item709',)=yes ('Item994',)=yes 171 ==> ('Item592',)=yes 104
18. ('Item961',)=yes 242 ==> ('Item592',)=yes 147    <conf:(0.61)>
19. ('Item43',)=yes 224 ==> ('Item592',)=yes 136    <conf:(0.61)> l
20. ('Item575',)=yes 175 ==> ('Item592',)=yes 106    <conf:(0.61)>
21. ('Item946',)=yes 223 ==> ('Item592',)=yes 135    <conf:(0.61)>
22. ('Item553',)=yes ('Item800',)=yes 182 ==> ('Item592',)=yes 110
23. ('Item803',)=yes 222 ==> ('Item592',)=yes 134    <conf:(0.6)> l
24. ('Item132',)=yes ('Item553',)=yes 174 ==> ('Item592',)=yes 105
25. ('Item732',)=yes 166 ==> ('Item592',)=yes 100    <conf:(0.6)> l
26. ('Item800',)=yes 392 ==> ('Item592',)=yes 236    <conf:(0.6)> l
27. ('Item676',)=yes 175 ==> ('Item592',)=yes 105    <conf:(0.6)> l
```

```
(624,)====>(592,)confidence:0.63
(495,)====>(592,)confidence:0.62
(578,)====>(592,)confidence:0.66
(732,)====>(592,)confidence:0.60
(575,)====>(592,)confidence:0.61
(676,)====>(592,)confidence:0.60
(870,)====>(592,)confidence:0.62
(828,)====>(592,)confidence:0.65
(501,)====>(592,)confidence:0.61
(737,)====>(592,)confidence:0.61
(613,)====>(592,)confidence:0.62
(803,)====>(592,)confidence:0.60
(946,)====>(592,)confidence:0.61
(43,)====>(592,)confidence:0.61
(907,)====>(592,)confidence:0.61
(456,)====>(592,)confidence:0.62
(863,)====>(592,)confidence:0.62
(961,)====>(592,)confidence:0.61
(994, 709)====>(592,)confidence:0.61
(571, 988)====>(709,)confidence:0.64
(553, 132)====>(592,)confidence:0.60
(132,)====>(592,)confidence:0.61
(432, 127)====>(592,)confidence:0.65
(571, 127)====>(709,)confidence:0.61
(800, 432)====>(592,)confidence:0.61
(800, 553)====>(592,)confidence:0.60
(800,)====>(592,)confidence:0.60
```

Identical association rules shows the result of this comparison.