Homework 2: Discovery of Frequent Itemsets and Association Rules

Team

Xiaoxu Gao < xiaoxu@kth.se >
Yu Wang < wang4@kth.se >

Solutions

- `ReadData` to read transaction data from file "`sdata.dat`", and processed the transaction line by line then stored it into `trans`.
- `findFrequentItemSets` to find frequent item sets in `trans` which make satisfied minimum support threshold.
  `findFrequentOneItemSets` to find frequent one item set. Based on frequent 1 item set, find frequent 2 item set,....., find L frequent item set until L-1 frequent item set does not exist.
  `aprioriGenCandidates` use join and prune to generate L candidate set based on L-1 frequent set.
- `genRule` to generate the association rules found in the data that satisfied the specified support and confidence

Files

- `Apriori.java` Algorithm to Find frequent itemsets and generate association rules
- `ItemSet.java` Extend TreeSet to store the data
- `ReadData.java` Handles data read
- `Rule.java` To generate the association rules
- `sdata.dat` dataset used, with one line per transaction and items seperated with space " "
- `sdata2.dat` another dataset used
- `result.txt` result of the dataset "sdata2.dat" with minimum support 3 and minimum confidence 0.6

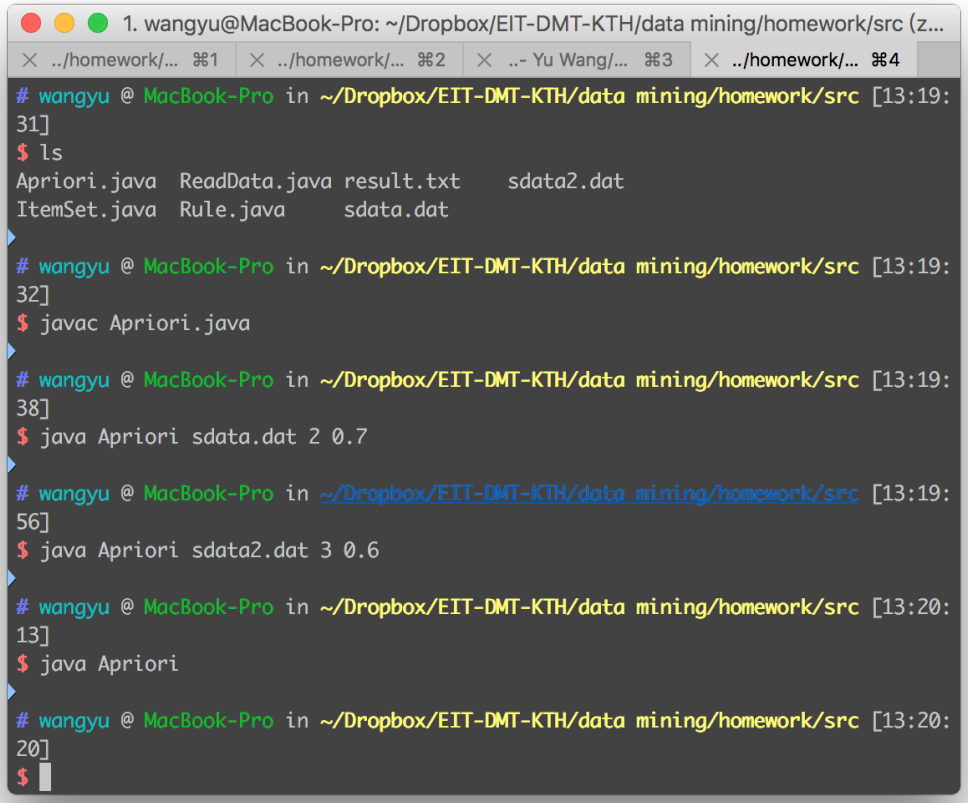
(Files should be in the same folder)

Build and Run

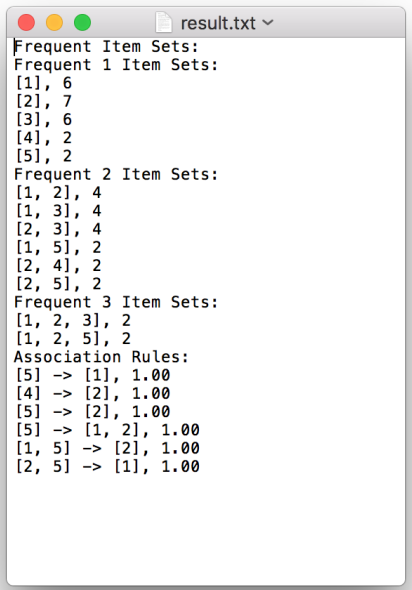
```
>javac Apriori.java
```

`>java Apriori <data_file> <min_sup> <min_conf>`, such as `java Apriori sdata.dat 3 0.6` , or just `java Apriori` to use default value.

Parameters, i.e., minimum support, minimum confidence and data file, could be changed in `Apriori.java`. And default values are min_sup = 2; min_conf = 0.7; data = "sdata.dat". It can also be specified at terminal as shown above. Example are shown below:


1. wangyu@MacBook-Pro: ~/Dropbox/EIT-DMT-KTH/data mining/homework/src (z...
../homework/... ⌘1
../homework/... ⌘2
..- Yu Wang/... ⌘3
../homework/... ⌘4
# wangyu @ MacBook-Pro in ~/Dropbox/EIT-DMT-KTH/data mining/homework/src [13:19:31]
$ ls
Apriori.java ReadData.java result.txt sdata2.dat
ItemSet.java Rule.java sdata.dat
# wangyu @ MacBook-Pro in ~/Dropbox/EIT-DMT-KTH/data mining/homework/src [13:19:32]
$ javac Apriori.java
# wangyu @ MacBook-Pro in ~/Dropbox/EIT-DMT-KTH/data mining/homework/src [13:19:38]
$ java Apriori sdata.dat 2 0.7
# wangyu @ MacBook-Pro in ~/Dropbox/EIT-DMT-KTH/data mining/homework/src [13:19:56]
$ java Apriori sdata2.dat 3 0.6
# wangyu @ MacBook-Pro in ~/Dropbox/EIT-DMT-KTH/data mining/homework/src [13:20:13]
$ java Apriori
# wangyu @ MacBook-Pro in ~/Dropbox/EIT-DMT-KTH/data mining/homework/src [13:20:20]
$

Result


result.txt
Frequent Item Sets:
Frequent 1 Item Sets:
[1], 6
[2], 7
[3], 6
[4], 2
[5], 2
Frequent 2 Item Sets:
[1, 2], 4
[1, 3], 4
[2, 3], 4
[1, 5], 2
[2, 4], 2
[2, 5], 2
Frequent 3 Item Sets:
[1, 2, 3], 2
[1, 2, 5], 2
Association Rules:
[5] -> [1], 1.00
[4] -> [2], 1.00
[5] -> [2], 1.00
[5] -> [1, 2], 1.00
[1, 5] -> [2], 1.00
[2, 5] -> [1], 1.00

The Result is stored in file `result.txt`, which includes Frequent Item Sets and Association Rules. Frequent Item Sets are represented like “[2, 5], 2”, item sets in bracket pairs followed by the corresponding support for the set. Association Rules are represented like “[5] -> [1, 2], 1.00”