Data Mining
Homework Assignment #2

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

February 20, 2014

CustomerID	TransactionID	BasketContent
1	1234	{Aspirin, Panadol}
1	4234	{Aspirin, Sudafed}
2	9373	{Tylenol, Cepacol}
2	9843	{Aspirin, Vitamin C, Sudafed}
3	2941	{Tylenol, Cepacol}
3	2753	{Aspirin, Cepacol}
4	9643	{Aspirin, Vitamin C}
4	9691	{Aspirin, Ibuprofen, Panadol}
5	5313	{Panadol, Vitamin C}
5	1003	{Tylenol, Cepacol, Ibuprofen}
6	5636	{Tylenol, Panadol, Cepacol}
6	3478	{Panadol, Sudafed, Ibuprofen}

Task 1

a. Compute the support and support count for itemsets {Aspirin}, {Tylenol, Cepacol}, {Aspirin, Ibuprofen, Panadol} by treating each transaction ID as a market basket.

b. Compute the confidence for the following association rules: {Aspirin, Vitamin C $\rightarrow$ Sudafed}, {Aspirin $\rightarrow$ Vitamin C}, {Vitamin C $\rightarrow$ Aspirin}. Why the results for last two rules are different?

c. List all the frequent itemsets under the support count threshold $s_{min} = 3$.

d. What does the anti-monotonicity property of a support imply? Give an example using the above data set.

Task 2

Write down all the steps of Apriori algorithm on the above data set under the support count threshold $s_{min} > 3$. How many steps of Apriori algorithm you

needed to perform? Draw a diagram showing all possible combinations of the items (e.g. lecture slide number 68). Mark all maximal, closed and infrequent items on this diagram.

Task 3

Construct an FP-tree using data set from Task 1 (use support count threshold $s_{min} > 3$) . Explain all the steps of the tree construction and draw a resulting tree. Based on this tree answer the questions: how many transactions contain {Aspirin} and {Cepacol}? How many transactions were made in total?

Task 4

Simulate frequent pattern enumeration based on the FP-tree constructed in the previous exercise. Report all the frequent patterns.

Task 5

Consider an example of a recommendation engine e.g. Amazon online shop, make an educated guess on how this kind of system is built. Find several weak points in Amazon recommendation engine that you would like to improve, explain.

Task 6 (2pt)

What is the probability to get 9 or 10 heads when you throw a fair coin 10 times? What is the probability to get 70 or more heads when you throw a fair coin 100 times? Conduct a computational experiment by generating 10,000 times such sequences of 10 coin tosses or 100 coin tosses.