Data Mining Homework Set 3, 2020

Cursus: BETA-INFOMDM Data Mining (INFOMDM)

Aantal vragen: 4

Data Mining Homework Set 3, 2020

Cursus: Data Mining (INFOMDM)

This is homework exercise set 3 of Data Mining

Aantal vragen: 4

1 Given are the following six transactions on items {A,B,C,D,E,F}:

tid	items
1	AB
2	AD
3	BCD
4	ACD
5	ACDF
6	ABE

Use the Apriori algorithm to compute all frequent item sets, and their support, with minimum support 2. For each level, make a table with the candidate frequent item sets, their support, and a check mark if the item set is frequent. Don't list as candidates item sets that do not need to be counted on the database. To generate the candidates, use the alphabetical order on the items.

Answer the following questions:

a. .... (0,5 pt.)

The support of item set AD is:

b. .... (0,5 pt.)

The number of level-2 candidates is:

c. .... (0,5 pt.)

The number of level-3 candidates is:

d. .... (0,5 pt.)

The number of frequent item sets is:

2 Given are the following five transactions on items {A,B,C,D,E}:

tid	items
1	ABC
2	ABC
3	BCD
4	BCD
5	CDE

Use the A-close algorithm to compute all closed frequent item sets, and their support, with minimum support 2. Use the alphabetical order on the items to generate candidates.

Answer the following questions:

a. .... (0,5 pt.)
The number of level-2 candidate generators is:

b. .... (0,5 pt.)
The number of level-3 candidate generators is:

c. .... (0,5 pt.)
The total number of generators is:

d. .... (0,5 pt.)
The number of closed frequent item sets is:

3 Consider the following database of sequences, containing orders in which different people have watched movies with Rocky Balboa as the main character:

id	viewing sequence
1	ROKR
2	ROCY
3	OCKR
4	YROKR
5	RORK

(R = Rocky, O = Rocky II, ... , Y = Rocky V).

Use the GSP algorithm to find all frequent sequences with minimum support of 3.

Answer the following questions:

The number of level-2 candidate frequent sequences is: **a.** ..(1 pt.)

The number of level-3 candidate frequent sequences is: **b.** ..(1 pt.)

The number of level-4 candidate frequent sequences is: **c.** ..(1 pt.)

The total number of frequent sequences is: **d.** ..(1 pt.)

4 In frequent tree mining, let T_1 = ab↑c and T_2 = abb↑a↑↑cb↑↑ac↑a be labeled rooted ordered trees.

Here we use the following string representation of an ordered labeled tree: list the labels according to the depth first pre-order traversal of the tree, and use the special symbol ↑ to indicate we go up one level in the tree.

Answer the following questions:

How many times does T_1 occur as an induced subtree of T_2? **a.** ..(1 pt.)

How many times does T_1 occur as an embedded subtree of T_2? **b.** ..(1 pt.)

Thank you, goodbye!