NAME	**NIA**	**GRADE**

Mining of Massive Datasets (2019-2020)

SECOND MID-TERM (TT05, TT06, TT08)

WRITE YOUR ANSWERS CLEARLY IN THE BLANK SPACES. Please write clearly, as if you were trying to communicate something to another person who needs to understand what you write to be able to evaluate you properly. If an answer requires intermediate steps, please mark clearly your final response with a rectangle. If you answer with text, please underline the key words or phrases of your answer. If absolutely necessary, you can attach an extra sheet to your exam, indicating that the solution can be found in the extra sheet.

Problem 1 *2 points*

Suppose you are given 360 baskets as follows:

- Item 1 appears in all baskets.
- Item 2 appears in the first 180 baskets.
- Item 3 appears in the first 120 baskets.
- Item 4 appears in the first 90 baskets.
- Item 5 appears in the first 72 baskets.
- Item 6 appears in the first 60 baskets.

Answer the following questions, **providing a brief justification** of each answer.

1. What is the support of the itemset $\{3, 4, 5, 6\}$?

2. Is itemset $\{3, 4, 5, 6\}$ a closed itemset?

3. If the support threshold is 0.2778, which itemset or itemsets are frequent? (Answer is valid if you give all of them.)

4. If the support threshold is 0.4167, which itemset or itemsets are maximal? (Answer is valid if you give all of them.)

Problem 2 *3 points*

Suppose you are given 120 baskets as follows:

- Item 1 appears in all baskets: baskets number 1, 2, 3, 4, 5, 6, ..., 120
- Item 2 appears in 60 baskets: baskets number 2, 4, 6, 8, 10, 12, ..., 120
- Item 3 appears in 40 baskets: baskets number 3, 6, 9, 12, 15, 18, ..., 120
- Item 4 appears in 30 baskets: baskets number 4, 8, 12, 16, 20, 24, ..., 120

Answer the following questions, **providing a brief justification** of each answer, and remembering to express the support as a **relative frequency**:

1. What is the support of itemset $\{2, 3\}$?

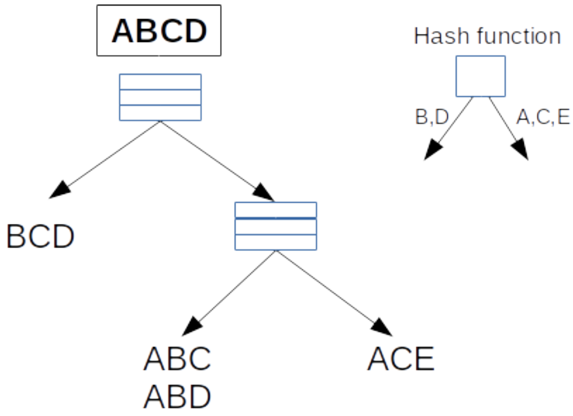
2. What is the confidence of the rule $\{2, 3\} \Rightarrow \{4\}$?

3. What is the lift of the rule $\{1, 2, 3\} \Rightarrow \{4\}$?

Tip: write the list of the 30 baskets in which item 4 appears, and circle in that list the baskets in which item 3 also appears.

Problem 3 *2 points*

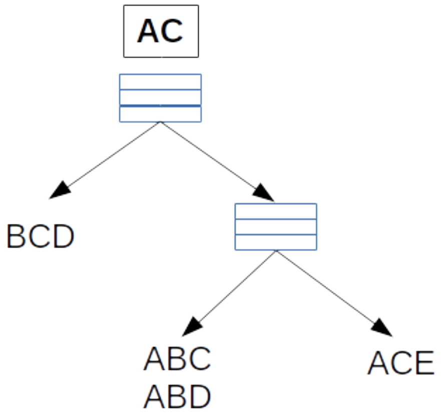
Consider the hash tree on the figure with its corresponding hash function, also depicted, in which 4 itemsets have been stored.



1. How many leaf nodes of this structure would be examined if you were searching for itemsets contained in the transaction $\{A, B, C, D\}$?

2. How many leaf nodes of this structure would be examined if you were searching for itemsets contained in the transaction $\{A, C\}$?

On the right, **draw the process** you used to find those leaf nodes, similarly to how we did it in class (as it appears in the theory materials).



Problem 4 *3 points*

Consider the following user ratings matrix for users 1, 2, 3 on items A, B, C, D. We are going to produce recommendations for user 2 using a user-based similarity method. An empty rating means the user has not seen the item yet.

	A	B	C	D
u_1	-1	$+1$	-1	$+1$
u_2	$+1$	-1		
u_3	$+1$	-1	$+1$	-1

1. Compute the similarity between user 1 and user 2, and between user 2 and user 3, using the similarity formula (*). Remember to mark your final answers with a rectangle.

2. Compute the score of item C for user 2, and of item D for user 2, using the scoring formula (**). Remember to mark your final answers with a rectangle.

3. Explain briefly what this method is doing, in your own words.

(*) Similarity formula for users u and v. $I_{u,v}$ is the set of items that both users have rated, u_i and v_i are the ratings of users u and v on item i, and $\hat{u}$ and $\hat{v}$ are the average ratings given by users u and v:

$$\mathrm{sim}(u,v) = \frac{\sum_{i \in I_{u,v}} (u_i - \hat{u}) \cdot (v_i - \hat{v})}{\sqrt{\sum_{i \in I_{u,v}} (u_i - \hat{u})^2 \cdot \sum_{i \in I_{u,v}} (v_i - \hat{v})^2}}$$

(**) Item scoring formula for item i for user u (note the absolute value in the denominator):

$$\mathrm{score}(u,i) = \hat{u} + \frac{\sum_{v:v_i \neq \mathrm{NULL}} \mathrm{sim}(v,u) \cdot (v_i - \hat{v})}{\sum_{v:I_{u,v} \neq \emptyset} |\mathrm{sim}(v,u)|}$$