NAME	NIA	GRADE

Mining of Massive Datasets (2023-2024)

MID-TERM EXAM

WRITE YOUR ANSWERS CLEARLY IN THE BLANK SPACES. Please write as if you were trying to communicate something in written to another person who is going to evaluate what you write. If for some reason (for example, if after you have written the solution you realize that there is some mistake that you would like to correct) you can attach an extra sheet to your exam. In this case, indicate clearly that the solution can be found in the extra sheet. Also, you may use other sheets to perform your calculations.

Problem 1 *1 point*

1. How do we call the process of determining if an item X in dataset 1 is equal to item Y in dataset 2?

Answer:

2. Define precisely what "data cleaning" is?

Answer:

Problem 2 *0.5 points*

If an attribute has values $5, 25, 50, 1000$ and you want to discretize it into a categorical attribute with two values "low" and "high" using EQUI-DEPTH ranges, in which category will each of these go?

- 5:
- 25:
- 50:
- 1000:

Problem 3 *0.75 points*

What is the Jaccard DISTANCE between vectors x = (1, 4, 5, 0, 0) and y = (0, 1, 4, 0, 5)?

Distance:

Problem 4 *0.75 points*

What is the Tanimoto SIMILARITY between vectors x = (1, 4, 5, 0, 0) and y = (0, 1, 4, 0, 5)?

Distance:

Problem 5 *4 points*

Consider the following documents.

- D1: "hello this is a test"
- D2: "goodbye this is a test"
- D3: "hello this is nice"
- D4: "testing this is nice"

1. Enumerate all six distinct shingles in this dataset, indicating their number (start from 1) and the text of the shingle. Use word trigrams as shingles. After, write a binary document/shingle matrix in which each column is a document and each row is a shingle.

   - Shingle 1=
   - Shingle 2=
   - Shingle 3=
   - Shingle 4=
   - Shingle 5=
   - Shingle 6=

2. Indicate the similarity between all pairs of documents, using the document/shingle matrix (Di is the column corresponding to document i).

   - Sim(D_1, D_2) =
   - Sim(D_1, D_3) =
   - Sim(D_1, D_4) =
   - Sim(D_2, D_3) =
   - Sim(D_2, D_4) =
   - Sim(D_3, D_4) =

3. Considering the following permutations: $\pi 1$ = (2, 6, 1, 4, 3, 5), $\pi 2$ = (5,2,3,6,1,4), and $\pi 3$ = (4,5,2,1,6,3), write the document/signature matrix in which each row is a permutation and each column is a document.

4. Indicate the similarity between all pairs of documents, using the document/signature matrix (Si is the signature of document i).

   - Sim(S_1, S_2) =
   - Sim(S_1, S_3) =
   - Sim(S_1, S_4) =
   - Sim(S_2, S_3) =
   - Sim(S_2, S_4) =
   - Sim(S_3, S_4) =

Problem 6 *3 points*

Consider the following database of transactions and minsup=0.3.

t1	a, b, c
t2	a, c, d
t3	b, c, d
t4	a, b, d, e
t5	b, c, e
t6	a, b, e

1. Indicate all 1-itemsets and their support, marking with an X the itemsets that do not satisfy the minsup criteria:

2. Execute a priori to generate all 2-itemsets and indicate their support, also marking with an X the itemsets that do not satisfy the minsup criteria:

3. Execute a priori to generate all 3-itemsets and indicate their support, also marking with an X the itemsets that do not satisfy the minsup criteria:

4. Indicate two rules that satisfy the minsup criteria, indicating the confidence of each rule: