Department of Computer Science

Prof. Dr. Nikolaus Augsten
Jakob-Haringer-Str. 2
5020 Salzburg, Austria
Telefon: +43 662 8044 6347
E-Mail: nikolaus.augsten@plus.ac.at

Paris Lodron Universität Salzburg

Similarity Search in Large Databases — Exam
Wintersemester 2023/2024 — 28.02.2024

Name: ____________________ **Student ID:** ____________

Hints

- Check whether you received all pages of the exam (9 pages).
- Write your name or your student ID on each sheet of the exam and hand in all pages.
- All answers are expected to be written on the exam sheets.
- Clearly highlight and enumerate additional pages that are used for longer answers. Match your text with the according exercise.
- Only use pencils that are permanent and non-red colored.
- Use the notation and techniques discussed in the lecture.
- Exercises with more than one solution are not graded.
- You are allowed to use one A4 sheet with your personal notes (both sides, hand written or printed).
- Exam duration: 90 minutes

Signature ____________________

Grading — Filled by the examiner

Exercise	1	2	3	4	5	6	7	8	Summe
Total points	2	2	2	2	2	2	2	2	16
Points reached									

Exercise 1 - *Blocking*. 2 Points

Consider the dirty entity resolution problem in Figure 1. Use the blocking technique on attribute ZIP to produce *candidate pairs*, i.e., the record pairs that must be compared. Illustrate the resulting blocks and list the candidates by their ID pairs.

Figure 1: Dirty entity resolution problem.

ID	Name	ZIP	YoB
b_1	Gruber	5034	1998
b_2	Smyth	5020	1993
b_3	Huber	5034	1949
b_4	Gruber	5020	2011
b_5	Chirsten	5020	1998
b_6	Huber	5034	1993

Exercise 2 - *String Edit Distance Algorithm.* 2 Points

Consider the given brute force string edit distance algorithm (cf. Algorithm 1) and perform the following tasks:

a) Draw the recursion tree for input strings *no* and *go*.

b) State the runtime complexity of Algorithm 1.

function **ed-bf(x, y)**
$m \leftarrow |x|$
$n \leftarrow |y|$
if $m = 0$ **then** **return** n
if $n = 0$ **then** **return** m
if $x[m] = y[n]$ **then** $c = 0$
else $c = 1$
return $\min(\text{ed-bf}(x, y[1 \ldots n-1]) + 1, \text{ed-bf}(x[1 \ldots m-1], y) + 1, \text{ed-bf}(x[1 \ldots m-1], y[1 \ldots n-1]) + c)$

Algorithm 1: Brute force string edit distance algorithm.

Exercise 3 - *q-Gram Distance*. 2 Points

Given the strings $x =$ `clapton` and $y =$ `chapman`. Compute the q-gram distance and the normalized q-gram distance between x and y ($q = 3$).

Exercise 4 - *Traversal Strings Lower Bound.* 2 Points

Compute the traversal string lower bound for the tree edit distance between trees T_1 and T_2 in Figure 2.

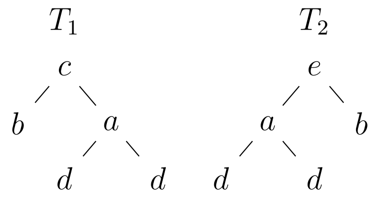

Figure 2: Two ordered trees T_1 and T_2.

Exercise 5 - *Forest Distance Matrix.* 2 Points

Consider ordered trees T_1 and T_2 in Figure 3, forest distance matrix fd, and tree distance matrix td for the trees T_1 and T_2.

Compute the values for the four shaded cells in the forest distance matrix fd.

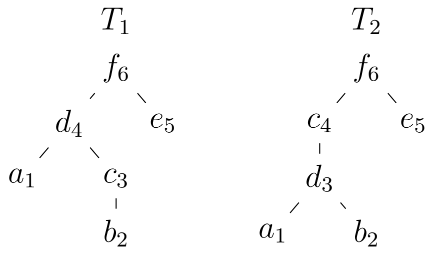

Figure 3: Two ordered trees T_1 and T_2.

fd:

	$d_j \rightarrow$	1	2	3	4	5	6
$d_i \downarrow$	0	1	2	3	4	5	6
1	1	0	1	2	3	4	5
2	2	1	0	1	2	3	4
3	3	2	1				
4	4	3	2				
5							
6							

td:

	1	2	3	4	5	6
1		1			1	
2	1	0	2	3	1	5
3	2	1	2	2	2	4
4		3			4	
5	1	1	3	4	0	5
6		5			5	

Exercise 6 - *Binary Branch Lower Bound.* 2 Points

Prove that the binary branch distance is a lower bound for the tree edit distance:

> *Let T_1 and T_2 be two trees. If the tree edit distance between T_1 and T_2 is $\delta_t(T_1, T_2)$, then the binary branch distance between them satisfies*
>
> $$\delta_{bb}(T_1, T_2) \leq 5 \times \delta_{ted}(T_1, T_2).$$

Exercise 7 - *Constrained Tree Edit Distance.* 2 Points

Consider ordered trees T_1 and T_2 in Figure 4. Compute the constraint tree edit distance and illustrate the according edit mapping between T_1 and T_2.

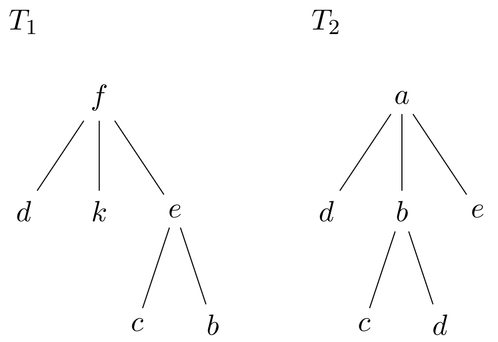

Figure 4: Two ordered trees T_1 and T_2.

Exercise 8 - *Dice Prefix Signature.* 2 Points

Consider the collection $R = \{s_1, s_2, s_3, s_4\}$ of sets in Figure 5. Compute *prefix signatures* for all sets $s_i \in R$ for *Dice similarity* threshold $t = 0.8$.

Note: For the Dice similarity, $Dice(r, s)$, between two sets, r and s, the following holds:

$$Dice(r, s) \geq t \Rightarrow |r \cap s| \geq \frac{t \cdot |r|}{2-t}$$

$$\begin{aligned} s_1 &= \{X, C, M, Z, F, N\} \\ s_2 &= \{Z, N, F, M, X\} \\ s_3 &= \{M, Z, F, G\} \\ s_4 &= \{C, M, G\} \end{aligned}$$

Figure 5: Set collection $R = \{s_1, s_2, s_3, s_4\}$.