Department of Computer Science

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

PARIS LODRON UNIVERSITÄT SALZBURG

Similarity Search in Large Databases — Exam
Winter Semester 2025/26 — 05.11.2025

Name: _____ **Student ID:** _____

Hints

- Check whether you received all pages of the exam (4 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 (single-sides, handwritten or printed).
- Exam duration: **45 minutes**

Signature _____

Grading — Filled by the examiner

Exercise	1	2	3	Sum
Total points	10	10	10	30
Points reached				

Exercise 1 - *Matching Techniques*. 10 Points

Consider the distance matrix in Figure 1. Two objects a, b are considered *mutual k-nearest-neighbors* if (1) a is one of the k closest neighbors of b and (2) b is one of the k closest neighbors of a. Compute a matching using the following techniques:

(a) k-nearest-neighbor matching with $k = 2$ from A to B,

(b) mutual k-nearest-neighbor matching with $k = 2$.

For each of the matching techniques, state whether they are symmetric in general.

Figure 1: Distance matrix, $A = \{a_1, a_2, a_3, a_4\}$, $B = \{b_1, b_2, b_3, b_4, b_5\}$.

	b_1	b_2	b_3	b_4	b_5
a_1	0.10	0.60	0.50	0.05	0.70
a_2	0.70	0.60	0.45	0.05	0.80
a_3	0.50	0.08	0.40	0.25	0.60
a_4	0.10	0.20	0.05	0.60	0.30

Exercise 2 - *Entity Resolution with Blocking*. 10 Points

Consider the clean-clean entity resolution problem in Figure 1. Use blocking to compute candidate pairs, i.e., the record pairs that must be compared. The blocking key consists of the 3-prefix (i.e., the first three letters) of the Name attribute.

(1) How many record pairs must be compared by the brute force approach?

(2) List all *candidate pairs* resulting from 3-prefix blocking on the Name attribute.

(3) Which candidate pairs are false negatives, which are false positives?

Table 1: Clean-clean entity resolution problem.

ID	Name	ZIP	YoB
a_1	Smith	5020	1993
a_2	Christen	5020	1998
a_3	Huber	5034	1993
a_4	Buber	5034	1949
a_5	Gruber	5010	2011

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

Solution of ER task: $\{(a_1, b_2), (a_2, b_5), (a_3, b_6), (a_4, b_3), (a_5, b_4)\}$

Exercise 3 - *String Edit Distance*. 10 Points

Compute the edit distance between the strings `'kitten'` and `'sitting'`. Show the matrix produced by the dynamic programming algorithm. Where, if anywhere, can you find the edit distance between a prefix of `'kitten'` and a prefix of `'sitting'` in the matrix? Where, if anywhere, can you find the edit distance between a suffix of `'kitten'` and a suffix of `'sitting'` in the matrix?