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
Winter Semester 2025/26

Exam
10.12.2025

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 - *Count Filter.* 10 Points

Consider the following query on the tables A and B in Figure 1 and the respective q-gram tables $QA(id, qgram)$ and $QB(id, qgram)$ $(k = 1, q = 2)$.

```
SELECT   A.id, B.id
FROM     A, QA, B, QB
WHERE    A.id = QA.id AND
         B.id = QB.id AND
         QA.qgram = QB.qgram AND
         ABS(LENGTH(A.name)-LENGTH(B.name)) <= k
```

(a) Show the q-gram tables QA, QB, and the query result.

(b) Extend the query with the count filter and show the resulting candidate pairs $(A.id, B.id)$. Although the query must remain general and not be tailored to the example, you may assume that kq is smaller than any of the q-gram profiles.

A

id	*name*
1	Jim
2	Lian

B

id	*name*
6	Jam
9	Liam

Figure 1: Two tables with names.

Exercise 2 - *Partition Signature.* 10 Points

Consider joining the sets in Figures 2 using Hamming threshold $h = 2$. Construct an appropriate partitioning and show the inverted list index for the partition signatures. Provide the set of candidates C and the join result M.

$$\begin{aligned} s_1 &= \{A, B, C, E\} \\ s_2 &= \{A, G, H\} \\ s_3 &= \{A, C, E\} \\ s_4 &= \{C, G, H\} \end{aligned}$$

Figure 2: Set collection $R = \{s_1, s_2, s_3, s_4\}$.

Partitioning $P =$

Partition Signatures:
$Par(s_1) =$
$Par(s_2) =$
$Par(s_3) =$
$Par(s_4) =$

Candidates $C =$

Join result $M =$

Inverted list index:

Exercise 3 - *Dice Prefix Signature*. 10 Points

Consider the collection $R = \{r_1, r_2, r_3, r_4\}$ of sets in Figure 3. Compute *prefix signatures* for all sets $r_i \in R$ for *Dice similarity* threshold $t = 0.8$.

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

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

$$\begin{aligned} r_1 &= \{Z, F, N, X, M, C\} \\ r_2 &= \{M, X, Z, N, F\} \\ r_3 &= \{Z, F, G, M\} \\ r_4 &= \{C, M, G\} \end{aligned}$$

Figure 3: Set collection $R = \{r_1, r_2, r_3, r_4\}$.