Fachbereich Informatik

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 — Prüfung
Wintersemester 2022/2023 — 17.03.2023

**Name:** _____ **Matrikelnummer:** _____

## Hinweise

- Bitte überprüfen Sie die Vollständigkeit des Prüfungsbogens (9 nummerierte Seiten).
- Schreiben Sie Ihren Namen und Ihre Matrikelnummer auf jedes Blatt des Prüfungsbogens und geben Sie alle Blätter ab.
- Grundsätzlich sollten Sie alle Antworten auf den Prüfungsbogen schreiben.
- Sollten Sie mehr Platz für eine Antwort benötigen, bitte einen klaren Verweis neben die Frage auf die Seitennummer des zusätzlichen Blattes setzen.
- Keinen Bleistift verwenden. Keinen roten Stift verwenden.
- Verwenden Sie die Notation und die Lösungsansätze, die während der Vorlesung besprochen wurden.
- Aufgaben mit mehr als einer Lösung werden nicht bewertet.
- Als Unterlage ist ein beliebig (auch beidseitig) beschriftetes A4-Blatt erlaubt.
- Zeit für die Prüfung: 90 Minuten

**Unterschrift** _____

## Korrekturabschnitt — Bitte frei lassen

| Aufgabe | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Summe |
|---|---|---|---|---|---|---|---|---|---|
| Maximale Punkte | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 16 |
| Erreichte Punkte | | | | | | | | | |

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 - *Longest Common Subsequence Distance.* 2 Points

Compute the longest common subsequence distance (LCS) between the strings `blank` and `blink`. Use the matrix produced by the dynamic programming algorithm to derive the shortest edit scripts (for the string edit distance) and represent them with the gap representation.

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

Given the strings $x =$ `carlson` and $y =$ `karlo`. 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

Prove that the preorder traversal string is a lower bound for the tree edit distance:

> *If two trees are at tree edit distance $k$, then the string edit distance between their preorder traversal strings is at most $k$.*

Exercise 5 - *Forest Distance Matrix.* 2 Points

Consider ordered trees $T_1$ and $T_2$ in Figure 2, 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 $fd$.

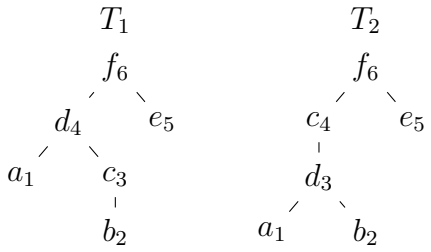


Figure 2: Two ordered trees $T_1$ and $T_2$.

$fd$:

| $d_i \downarrow$ / $d_j \rightarrow$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 0 | 1 | 2 | 3 | 4 | 5 |
| 2 | 1 | 0 | 1 | 2 | 3 | 4 |
| 3 | 2 | 1 | | | | |
| 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 - *Traversal Lower Bound.* 2 Points

Consider ordered trees $T_1$ and $T_2$ in Figure 3. Compute the traversal lower bound (TLB) between $T_1$ and $T_2$ for pre- *and* postorder.

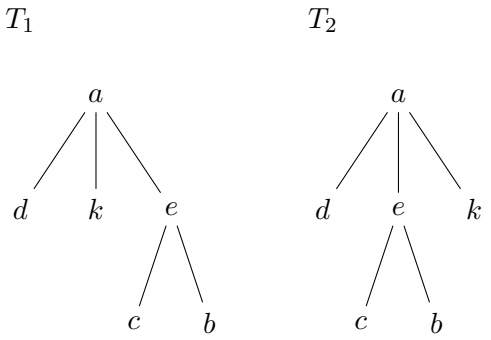


Figure 3: Two ordered trees $T_1$ and $T_2$.

Exercise 7 - *Binary Branch Distance and Lower Bound.* 2 Points

For the ordered trees $T_1$ and $T_2$ in Figure 4:

a) Represent $T_1$ and $T_2$ as normalized binary trees and compute the binary branch distance.

b) Based on the binary branch distance, what is the smallest value that the edit distance between $T_1$ and $T_2$ can adopt?

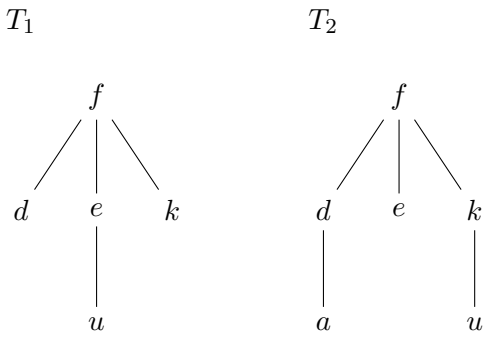


Figure 4: Two ordered trees $T_1$ and $T_2$.

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:

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

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

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