# Übungsprojekt C

# ETL: Deduplikation

Eine wichtige Aufgabe vor dem Laden von Daten im ETL Prozess ist das Erkennen von Duplikaten. Dies ist wichtig, wenn verschiedene Datenquellen geladen werden müssen, zum Beispiel Produktkataloge aus mehreren Verkaufsbereichen, um redundante Daten zu erkennen und zusammen zu führen. Aber auch bei einer einzigen Datenquelle können Duplikate vorliegen, wenn das Datenbanksystem keine Constraints prüft, was in Unternehmen durchaus Praxis ist.

In der Vorlesung haben Sie mehrere Distanzen kennengelernt, um die Gleichheit von Entitäten, zum Beispiel Produkten, zu messen. Mit Hilfe dieser Distanz lässt sich eine Ähnlichkeitsmatrix aufstellen, die wegen ihrer Größe aber mit in der Vorlesung vorgestellten Heuristiken nur teilweise untersucht werden kann.

Das Übungsprojekt besteht dabei aus den folgenden Teilaufgaben:

1. Schreiben Sie ein Java/C++/C Programm, das die beigefügte Datei mit Restaurantnamen einliest und mit Hilfe der Editier-Distanz eine vollständige Ähnlichkeitsmatrix erstellt.

2. Bestimmen Sie einen sinnvollen Schwellwert, um Duplikate unter den Namen zu identifizieren.

3. Ändern Sie Ihr Programm nun so, dass nicht mehr die vollständige Matrix berechnet wird, sondern eine Partition in $b$ Blöcke. Identifizieren Sie die Duplikate mit demselben Schwellwert wie in Punkt 2. Stellen Sie das Ergebnis in Abhängigkeit von der Blockgröße $b$ dar.

4. Durch welche Methoden können auch bei geringer Blockgröße relativ viele Duplikate erkannt werden? Beurteilen Sie Ihre Lösung im Hinblick auf diese Frage.