Universität Leipzig
Institut für Informatik
Abteilung Datenbanken

Dozent: Dr. Victor Christen
Abgabe: 08.06.2022

# Data Wrangling

Sommersemester 2022

—

## Praktische Übung: Blocking

## Vorbereitung

Für die Bearbeitung der Aufgaben sollten Sie den entsprechenden Zip-Ordner im Moodle-Kurs herunterladen. Dieser enthält das Code-Skelett und Datensätze für die Verwendung des Record-Linkage Systems. Die Module und die Datensätze sollten Sie in Ihrem Python-Projekt einbinden.

Vor der Implementierung sollten Sie sich mit den einzelnen Modulen und der Struktur für die unterschiedlichen Aufgaben vertraut machen. Diese Übung umfasst lediglich das Vervollständigen des `blocking.py` Moduls. Für das nähere Verständnis können Sie `recordLinkage.py` so wie es ist ausführen. Es verwendet die bereitgestellten Datensätze und bereits implementierte Funktionen. Somit ist die Ausgabe der unterschiedlichen Module nachvollziehbar. Für das Testen des Programms auf Korrektheit ist empfehlenswert die Datensätze ohne Korruption zu verwenden, die `clean-A-1000.csv` und `clean-B-1000.csv` heißen. Danach können die anderen Datensätze verwendet werden.

**Aufgabe 1: Soundex-Code** (2 Punkte)

Erarbeiten Sie sich wie der Algorithmus für Soundex-Blocking funktioniert und berechnen Sie manuell den Soundex für die folgenden Namen (Annahme: Konvertierung zu Kleinschreibung):

- arnold
- schwarzenegger
- Ihr Vorname (nur Ersten)
- Ihr Nachname (nur Ersten)

Bei der Beantwortung im Moodle müssen Sie die einzelnen Schritte für die Konvertierung darstellen, die den Schritten des Kapitels *Blocking* entsprechen. Es gibt pro richtige Antwort inklusive richtiger Schritte 0.5 Punkte. Die Antwort wird mit 0 Punkten bewertet, wenn die Zwischenschritte fehlen.

**Hinweis:** Es existieren verschiedene Implementierungen für die Berechnung von Soundex. Verwenden Sie bitte die Beschreibung aus dem Kapitel *Blocking*.

**Aufgabe 2: Soundex-Blocking** (3 Punkte)

Bei dieser Aufgabe soll der Soundex Algorithmus in dem Modul `blocking.py` implementiert werden. Die resultierenden Soundex-Codes sollen als Blocking-Key verwendet werden, dass heißt Records mit dem gleichen Soundex-Code bilden einen Block. Die Implementierung soll den Schritten aus der Vorlesung entsprechen.

(a) (1 Punkt) Generieren Sie für die folgenden Namen die Soundex-Codes: *christina*, *kirstyn*, *allyson*, *alisen*.

(b) (1 Punkt) Führen Sie `simpleBlocking` und `Soundex` basiertes Blocking für die Attribute aus, die Sie als passend erachten. Dabei sollen die Datensätze `clean-A-1000.csv` und `little-dirty-A-1000.csv` betrachtet werden. Dokumentieren Sie die Anzahl der Blöcke, die minimale, maximale und durchschnittliche Größe bei der Anwendung.

(c) (1 Punkt) Beschreiben Sie Ihre Ergebnisse in einigen Sätzen und begründen Sie die Wahl Ihrer Attribute.

Für die Frage 1 erhalten Sie 1 Punkt, wenn alle Codes korrekt sind. Andernfalls wird der Python-Code für die Bewertung mit verwendet. Für Frage 2 und 3 erhalten Sie die volle Punktzahl für jede realistische Antwort und Erklärung bzgl. der Wahl der Blocking-Attribute.

**Aufgabe 3: SLK-581** (3 Punkte)

Der *Statistical Linkage Key SLK-581* kann für die Identifikation von Records, die die gleiche Person repräsentieren verwendet werden. Die Generierung eines SLK-581 Keys wird im Kapitel *Record-Linkage - Blocking* behandelt. Ein detaillierte Beschreibung des Verfahrens finden Sie im Moodle-Kurs.

Testen Sie Ihre Implementierung an folgenden Personen-Records (Annahme: Vorverarbeitung Kleinschreibung).

1. john johnson, 19.05.1967
2. maria meier, 11.11.1911
3. al hu, 01.12.2012
4. yi zu, 01.10.2010

Sie erhalten 0.25 Punkte für jeden korrekten Key.

(a) (1 Punkt) Führen Sie wie beim Soundex eine Analyse der resultierenden Blöcke für `clean-A-1000.csv` und `little-dirty-A-1000.csv` Datensätze durch (Anzahl der generierten Blöcke sowie maximale,minimale und durchschnittliche Größe).

(b) (1 Punkt) SLK-581 wurde in Australien im Bereich des Gesundheitswesen entwickelt. Welche Aspekte wären problematisch, wenn es im deutschsprachigen Raum angewendet wird? Wie müsste das Verfahren modifiziert werden, um effektiver zu sein.

**Aufgabe 4: Bewertung der Performance** (2 Punkte)

Evaluieren Sie die implementierten Blocking-Verfahren aus Aufgabe 2 und 3 für die sauberen und die leicht verschmutzten Daten mit jeweils 1000 und 10000 Records. Beschreiben Sie Ihre Resultate. Bewerten Sie dabei die Resultate, bzgl. der Anzahl der Blöcke, Verteilung der Blockgrößen(Minimum, Maximum, Durchschnitt und Median) sowie die Laufzeiten für die verschiedenen Datensätze.

Ermitteln Sie ebenfalls wie viele Blocking-Keys gemeinsam auftreten und wie viele unterschiedlich sind zwischen 2 unterschiedlichen Datenquellen.

Im Moodle geben Sie bitte die numerischen Resultate im ersten Textfeld an und im zweiten Textfeld die Beschreibung und Erklärung der Resultate. Gehen Sie dabei auf folgende Punkte ein:

- Welche Blocking-Funktionen und Blocking-Keys würden Sie verwenden und weshalb?
- Welche Attribute sind isoliert nicht zum Blocking geeignet, führen aber in Kombination zu einer Verbesserung?
- Definieren Sie eine Menge von Kriterien, die ein guter Blocking-Schlüssel erfüllen sollte. Begründen Sie Ihre Antwort mithilfe der Resultate Ihrer Experimente.

Sie erhalten 1 Punkt pro Antwort, wobei die Vollständigkeit der numerischen Werte für den 1. Teil relevant sind und für den 2. Teil passende Begründungen und Erklärungen.