# Algorithmen und Datenstrukturen

## Suche in Texten

### Aufgabe 1 Levenshtein Distanz [2 Punkte]

Die Levenshtein Distanz gibt an, wie viele Editieroperationen zum Überführen des einen Strings in den andern notwendig sind. Bestimmen Sie die Levenshtein Distanz folgender String Paare.

```
AUSTAUSCH – AUFBAUSCH
BARBAREN – BARBARA
COCACOLA – COCAINA
```

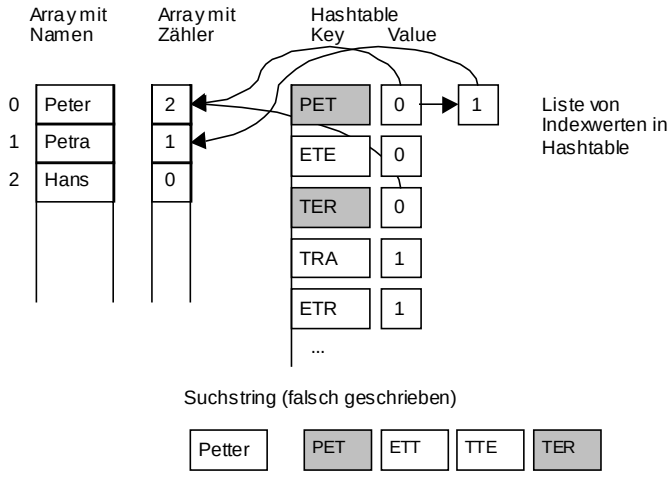
### Aufgabe 2 Reguläre Ausdrücke [3 Punkte]

a) Lösen Sie die Aufgabe in den Vorlesungsunterlagen.

b) Definieren Sie Reguläre Ausdrücke für:
Eine IP Adresse: 12.122.12.1 oder 198.168.1.1
Eine E-Mail Adresse: hans.muster@zhaw.ch

### Aufgabe 3 Fehlertolerante Suche [5 Punkte]

Es soll eine fehlertolerante Suche nach dem Namen (der Rangliste) implementiert werden. Für diesen Zweck wird die sogenannte Trigram Methode angewandt. Die Idee dabei ist, dass der String in 3-Buchstaben Gruppen unterteilt wird, z.B. sind das bei „Peter“, die 3-er Gruppen „PET“, “ETE“, “TER“. Diese 3-er Gruppen werden für alle vorkommenden Namen gebildet und wie in folgender Abbildung gezeigt abgespeichert.




Die 3-er Gruppen werden in eine Hashtabelle als Schlüssel abgelegt; der dazugehörige Wert entspricht einer Liste von (Array-)Indexwerten der entsprechenden Namen, z.B. PET kommt in Peter und Petra vor, demnach enthält die Liste 0 und 1 (Arrayindex von Peter und Petra).

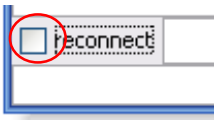
Bei der Suche wird der Suchstring ebenfalls in 3-er Gruppen unterteilt. Diese Gruppen werden nun in der Hashtabelle gesucht und der Zähler des entsprechenden

Namens um eins inkrementiert. Das wird für alle 3-er Gruppen des Suchstrings gemacht. Derjenige Namen mit dem höchsten Zählwert am Schluss ist dann der gesuchte.

Aufgabe: Implementieren Sie einen ***FuzzySearchServer*** dem Sie die Rangliste übergeben und einen Trigram Index erzeugen können.

Es soll dann nach einem Namen mit z.B. "FIND Kiptam Dani" gesucht werden können und es sollte "Kiptum Daniel" gefunden werden.

- Hinweis: es  ist Vorteil, wenn Sie eine Methode schreiben, die die Zerlegung durchführt und eine Liste der Trigrammen zurück gibt.
- Map<String,List> zur Speicherung der Trigramme verwenden.
- Ihr Server kann anhand des ersten Strings ("FIND") erkennen, ob es sich um die Eingabe einer Namensliste oder um eine Anfrage handelt.
- Damit die Werte zwischen Aufrufen in Ihrem Server gespeichert bleiben, müssen Sie "reconnect" disablen.




- Unter http://radar.zhaw.ch/adressbook/Adressbook.aspx finden Sie eine - vom Dozenten entwickelte und auch in einem Milliarden-Konzern eingesetzte - unscharfe Suche für Namen, die auf einer Kombination aus Trigram, Levenshtein und Phonetischer Suche aufbaut.