Algorithmen und Datenstrukturen

Suche in Texten

Aufgabe 1 Levenshtein-Distanz

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

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

Aufgabe 2 Reguläre Ausdrücke (Regex)

a) Definieren Sie reguläre Ausdrücke für eine IP-Adresse, z.B: 12.122.12.1 oder 198.168.1.1.

b) Definieren Sie reguläre Ausdrücke für eine E-Mail-Adresse, z.B.: hans.muster@zhaw.ch

Aufgabe 3 Suche nach Muster

Es soll auf Webseiten z.B. https://tel.search.ch/?was=<Name> nach Telefonnummern gesucht werden (der gesuchte Name wird in der ExBox eingegeben). Überlegen Sie sich das Regex-Muster für (Schweizer) Telefonnummern und suchen Sie auf obiger Seite alle Nummern. Zur Vereinfachung dürfen Sie davon ausgehen, dass die Telefonnummern ohne Ländervorwahl sind, 10 Ziffern lang sind und mit 0 beginnen. Kurznummern und Mehrwegdienste dürfen Sie ebenfalls vernachlässigen. Stellen Sie aber sicher, dass die Vorwahl korrekt ist. Beachten Sie auch, dass Leerzeichen optional sind (siehe auch https://de.wikipedia.org/wiki/Telefonvorwahl_(Schweiz)). Vorsicht beim Testen, die Homepage lädt beim Nach-unten-Scrollen automatisch weitere Daten (die fehlen dann in Ihrem Resultat).

Das Laden der Web Seite kann mit folgendem Code durchgeführt werden.

```
import java.net.*;
import java.io.*;

String inputLine;

URL oracle = new URL("http://www.oracle.com/");
BufferedReader in = new BufferedReader( new InputStreamReader(oracle.openStream()));
while ((inputLine = in.readLine()) != null) {
   System.out.println(inputLine);
}
in.close();
```

Hinweis zum Lesen der Homepage:

Siehe auch Beispielcode HttpsClient.java. Das Codebeispiel muss leicht ergänzt und angepasst werden.

Aufgabe 4 (optional)

Erweitern Sie Ihr Programm so, dass auch URLs erkannt werden und die Seiten dieser URL rekursiv gefolgt wird (bis zu einer vorbestimmten Tiefe). Achten Sie darauf, dass Sie das die Verlinkung auch Zyklen enthalten kann und es unterschiedliche Arten von Links gibt.