Hochschule Furtwangen
Fakultät Informatik
Prof. Dr. Lothar Piepmeyer

Praktikum ‚Datenbanken'

Aufgabenblatt 5

In diesen Aufgaben beschäftigen Sie sich mit Daten, die im Rahmen einer wissenschaftlichen Arbeit aus der Botanik gesammelt wurden. In vielen Fällen – und so auch im vorliegenden Fall – müssen die Daten bereinigt werden, bevor sie genutzt werden.

Aufgabe 1:
a. In der Datei `barcodes.csv` finden Sie Daten über DNA-Informationen einiger Pflanzen. Verschaffen Sie sich zunächst mit einer Spreadsheet-Software wie Excel einen Überblick über die Daten: Programme wie Excel haben eine Import-Möglichkeit für Dateien im CSV-Format (CSV steht für comma separated values). Versuchen Sie zumindest ansatzweise die Bedeutung der Daten zu erfassen. Bei der Arbeit mit Daten ist es wichtig, sich einen ersten Überblick zu verschaffen, bevor man sie in eine Datenbank importiert. Lassen Sie sich dabei von den folgenden Fragen leiten:

- Welche SQL-Datentypen erscheinen Ihnen für die Spalten geeignet? Gibt es Spalten, für die der Typ `int` genutzt werden kann?
- Gibt es Spalten, für die not-null- oder unique-Constraints passen würden?
- Gibt es Spalten mit nur wenigen verschiedenen Werten?
- Gibt es einen geeigneten Primärschlüssel

b. Da wir die Daten in eine SQL-Tabelle importieren wollen, sollte Sie die Länge der Texte in der Tabelle interessieren. Ermitteln Sie für jede Spalte mit Hilfe des Formel-Editors Ihrer Spreadsheet-Software die maximale Länge der Werte.

c. Ist die folgende Tabelle geeignet, um die Daten aufzunehmen? Korrigieren Sie die Tabelle falls nötig. Die Größenangaben in den `varchar`-Spalten dürfen aber nicht verschwenderisch sein!

```
create table barcodes (
    id int generated always as identity primary key,
    taxon varchar(20) not null,
    type varchar(1),
    voucherID varchar(12),
    gardenID varchar(12),
    sampleID varchar(12),
    origin varchar(50),
    rpoB varchar(12),
    rpoC varchar(12),
    matK varchar(12),
    trnHpsbA varchar(12),
    rpl32 varchar(12)
)
```

Hochschule Furtwangen
Fakultät Informatik
Prof. Dr. Lothar Piepmeyer

Aufgabe 2:
a. Importieren Sie die Daten aus der Datei in die Tabelle `barcodes`. Orientieren Sie sich dabei an der Aufgabe 4e aus Aufgabenblatt 2. Beachten Sie aber, dass es in der csv-Datei keine Spalte gibt, die zum Primärschlüssel der Tabelle gehört. Die insert-Anweisung bedarf daher besonderer Beachtung!

b. Prüfen Sie mit Hilfe einer select-Anweisung , ob wirklich alle Datensätze importiert wurden.

c. Gibt es eigentlich Spalten mit Dubletten in der Tabelle? Sie wissen bereits, dass die Anweisung `select distinct` Dubletten nicht anzeigt. Nutzen Sie diese Information.

Aufgabe 3:
a. Wie viele verschiedene Werte gibt es in der Spalte `taxon`?

b. Da die Spalte `taxon` viele Dubletten enthält, lagern wir dieses Werte in eine neue Tabelle aus:

```
create table taxa(
  id int generated always as identity primary key,
  name varchar(48),
)
```

Diese Tabelle enthält einen kleinen, aber gravierenden Design-Fehler. Finden und korrigieren Sie den Fehler.

c. Überlegen Sie, wie Sie die verschiedenen Werte der Spalte `taxon` in die Tabelle `taxa` kopieren können und setzen Sie Ihre Erkenntnisse in die Praxis um. Sie können sich an Aufgabe 4d in Aufgabenblatt 3 orientieren.

d. Prüfen Sie, ob wirklich alle Daten kopiert wurden. Stellen Sie auch sicher, dass keine überflüssigen Werte mehr vorkommen.

e. Jetzt sollen in der Tabelle `barcodes` die Werte der Spalte `taxon` nach und nach durch die IDs der Tabelle `taxa` ersetzt werden: Erweitern Sie die Tabelle `barcodes` um eine neue ganzzahlige Spalte `taxonid`. Welche Werte gibt es jetzt in der Spalte `taxonid`?

f. Die SQL-Anweisung `update` kennen Sie aus Aufgabenblatt 1. Was machen die folgenden Anweisungen?

```
create table t1(
     c11 int,
     c12 varchar(20)
)
create table t2(
     c21 int,
     c22 varchar(20)
)
```

```
-- fill data into t1 and t2
update t1
set c11=(select c21 from t2 where c12=c22);
```

g. Formulieren Sie eine SQL-Anweisung, um in die Spalte `taxonid` den passenden Wert aus der Tabelle `taxa` einzutragen. Sie können sich dazu an der update-Anweisung aus Aufgabenteil f. orientieren.

h. Machen Sie die Spalte `taxonid` zum Fremdschlüssel für die Tabelle `taxa`

i. Machen Sie sich klar, dass die Spalte `taxon` in der Tabelle `barcodes` keine Existenzberechtigung mehr hat. Löschen Sie die Spalte `taxon` aus der Tabelle `barcodes`.