Prof. Dr. Stefan Brass
PD Dr. Alexander Hinneburg
Institut für Informatik
MLU Halle-Wittenberg

14.10.2024

Einführung in Datenbanken
— Übungsblatt 1 (Web-Quellen, CSV-Format) —

Ihre Lösungen laden Sie bitte in die Übungsplattform in StudIP hoch ([StudIP-Eintrag der Vorlesung], Reiter „Übungsplattform“).

Einsendeschluss ist Montag, der 21.10.2024, 18^{00}.

Hausaufgaben können einzeln oder in Zweier-Gruppen bearbeitet werden. Sie können die Gruppe für jede Aufgabe neu wählen. Nutzen Sie die Möglichkeit zur Gruppen-Abgabe nur, wenn Sie so wirklich mehr lernen als bei Einzelabgabe! Wenn Sie die Gruppenabgabe nur als Arbeitsersparnis sehen, und die Aufgaben aufteilen, wird Ihnen das bei der Klausur auf die Füße fallen. Die Menge an Hausaufgaben ist für Einzelabgabe gedacht. Sie sollten sich vor dem Gruppentreffen allein mit jeder Aufgabe beschäftigt haben.

Denken Sie daran, dass Sie bei jeder Aufgabe angeben müssen, ob Sie bereit sind, vorzurechnen:

- „`VORRECHNEN:0`“: Ich werde nicht zur Übung kommen.
- „`VORRECHNEN:1`“: Ich möchte diese Aufgabe nicht vorrechnen.
- „`VORRECHNEN:2`“: Ich möchte diese Aufgabe nur ungern vorrechnen.
- „`VORRECHNEN:3`“: Ich kann vorrechnen, lasse aber gern anderen den Vortritt.
- „`VORRECHNEN:4`“: Ich kann problemlos vorrechnen.
- „`VORRECHNEN:5`“: Ich möchte gerne, dass meine Abgabe besprochen wird.

Falls Sie als Gruppe abgeben, muss jedes Gruppenmitglied einzeln die Bereitschaft zum Vorrechnen erklären (`VORRECHNEN1:N` ist der Wert für den Studierenden, der die Aufgabe in die Übungsplattform hochgeladen hat, und `VORRECHNEN2:M` der Wert für den anderen Studierenden).

„Zu ähnliche Lösungen“ (Plagiate) führen automatisch zu 0 Punkten für alle Beteiligten. Das gilt auch dann, wenn Sie nicht direkt abgeschrieben haben, sondern nur zufällig die gleiche Quelle benutzt haben (ohne diese zu nennen). Sie haben aber die Möglichkeit, Quellen offenzulegen. In diesem Fall zählt Ihre Abgabe nicht als Plagiat, sondern wird normal korrigiert. Im Kapitel 0 der Vorlesungsfolien wird erläutert, wie Sie die Angaben zu codieren haben.

Beispiele für Quellenangaben:

- „`ChatGPT:3`“ für eine von ChatGPT generierte Lösung (ohne manuelle Überarbeitung, sonst „`ChatGPT:2`“, bei Korrektur einer eigenen Lösung durch ChatGPT auch „`ChatGPT:1`“),
- „`STUD:3`“ für eine von einem anderen Studierenden direkt kopierte Lösung („`STUD:2`“ bei eigener Überarbeitung und „`STUD:1`“ für wesentliche fremde Hilfe zur Korrektur eines eigenen Lösungsversuches),
- „`ALT:3`“ für eine fremde Lösung aus einem früheren Semester (entsprechend auch mit Abstufungen), und
- „`QUELLE:` ⟨`Freitext`⟩“ für weitere Quellen.

Es versteht sich von selbst, dass wir nicht empfehlen, fremde Lösungen zu kopieren oder sich Lösungen von ChatGPT generieren zu lassen. Wieviel Sie lernen, hängt immer auch am eigenen Einsatz.

Wir haben ein Forschungsprojekt zur Erkennung von Plagiaten in SQL-Anfragen, und würden die so gewonnenen Daten in pseudonymisierter Form gern dafür nutzen. Dafür ist auch wichtig, dass Sie bitte **nie den eigenen Namen in die Abgaben schreiben**.

Umgekehrt müssen Sie damit rechnen, dass wir bisher erstellte Werkzeuge auf den Abgaben ausprobieren, um ggf. undeklarierte Kopien zu erkennen. Ersparen Sie sich und uns den Ärger.

Vielleicht hilft Ihnen auch diese Ehrlichkeit sich selbst gegenüber. Wenn Sie alle Abgaben kopieren, und immer „`STUD:3`“ in die Lösung schreiben, würden Sie die Studienleistung zwar bekommen (je nach Fähigkeit Ihres Kollegen), aber Sie haben auch für sich selbst dokumentiert, dass Sie dem für die Klausur notwendigen Wissen und damit den Leistungspunkten für diesen Kurs kein Stück näher gekommen sind.

In Kombination mit den Angaben zum Vorrechnen können Sie allerdings auch deutlich machen, dass Sie eine kopierte oder von ChatGPT generierte Lösung aber vollständig verstanden haben. Sie könnten daraus für sich zumindest mehr Wissen ziehen, als wenn Sie sich mit der Aufgabe gar nicht beschäftigen würden.

Wir empfehlen natürlich immer die selbständige Bearbeitung der Hausaufgaben.

Freiwillige Zusatzangabe für Forschungsprojekt:
Ich wäre Ihnen dankbar, wenn Sie zusätzlich die verwendete Arbeitszeit in Minuten für diese (und alle weiteren Hausaufgaben) angeben würden. Verwenden Sie bitte das Format „`ZEIT:N`“ (bei Gruppenarbeit „`ZEIT1:N`“ und „`ZEIT2:M`“). Wenn Sie z.B. eine Stunde für diese Hausaufgabe gearbeitet haben, schreiben Sie „`ZEIT:60`“. Falls mindestens 20 Studierende für jeweils mindestens die Hälfte der SQL-Aufgaben die Arbeitszeit angeben, wird unter diesen ein 50 € Amazon-Gutschein verlost (auf Wunsch auch als Bargeld). Bei deutlich mehr Studierenden wird die Anzahl Preise vermutlich auf 2 oder 3 erhöht. Diese Aufgabe zählt allerdings noch nicht mit, da es keine SQL-Aufgabe ist. Sie würden trotzdem auch einen Beitrag zur Verbesserung der Lehrveranstaltung leisten.

Aufgabe 1 (8 Punkte)

Es sollen die auf der Vorlesungs-Webseite verlinkten Internet-Quellen gemeinsam bewertet werden (wenn Sie wollen, können Sie auch weitere Quellen vorschlagen).

Es soll gemeinsam eine Tabelle der folgenden Art erstellt werden:

WEBSEITEN				
STUD	NR	URI	ART	BEWERTUNG
12345	0	VORRECHNEN:4		
12345	1	https://de.wikipedia.org/wiki/SQL	Lexikon-Artikel	+
12345	2	https://www.postgresql.org/	DBMS-Homepage	++

Es sollen später alle Abgaben in eine gemeinsame Tabelle eingefügt werden. Da aber nicht Ihr Name oder andere identifizierende Daten in den Abgaben stehen sollen, würfeln Sie bitte eine fünfstellige Zufallszahl für die erste Spalte aus (und verwenden dann die gleiche Zahl in allen Ihren Tabellenzeilen). Damit ist die Wahlscheinlichkeit, dass zwei verschiedene Studierende die gleiche Zahl verwenden (und es dann bei der Vereinigung aller Tabellen zu einem Fehler kommt), zumindest recht klein. Sie können sich die Zufallszahl in PostgreSQL mit folgender Anfrage generieren lassen:

```
SELECT CAST(floor(random()*100000) AS INTEGER)
```

Die Funktion `random()` liefert eine Zufallszahl x mit $0 \leq x < 1$. Die Funktion `floor(x)` rundet ab auf die nächste ganze Zahl, liefert das Ergebnis aber als Gleitkommazahl, so dass anschließend eine Typ-Umwandlung mit `CAST(`x `AS` T`)` in den Typ T durchgeführt werden muss (hier `INTEGER`). Die obige Anfrage können Sie in den Adminer eingeben (Passwort in Übung oder StudIP, Reiter „Datenbank“). Im Adminer müssen Sie links (oder unten) im Menu „`SQL command`“ auswählen. Dann bekommen Sie eine Dialogbox zur Eingabe einer SQL-Anweisung. Nachdem Sie die Anfrage eingegeben haben, drücken Sie auf den „`Execute`“-Knopf. Es ist erwünscht, dass Sie wirklich PostgreSQL für diese Berechnung verwenden.

Die Nummer in der Spalte `NR` wählen Sie fortlaufend, von 1 beginnend. Sie müssen eine Liste von fünf Webseiten abgeben. Den Wert `0` verwenden Sie für die Angaben zum Vorrechnen und ggf. für Quellenangaben sowie die verwendete Arbeitszeit (alles in die Spalte `URI`, durch Leerzeichen getrennt, die anderen beiden Spalten bleiben frei).

In die dritte Spalte schreiben Sie bitte die Webadresse (URI) Bevorzugt sollten die Webseiten aus der Linkliste zur Vorlesung stammen:

- [https://users.informatik.uni-halle.de/~brass/db24/links.html]
- [https://users.informatik.uni-halle.de/~brass/db24/software.html]

Sie dürfen aber auch weitere Webseiten wählen, die allerdings klar einen Bezug zu Datenbanken haben müssen. Wenn Sie nützliche Webseiten entdecken, die in der Link-Liste zur Vorlesung fehlen, schicken Sie dem Dozenten bitte eine Email.

In die vierte Spalte „`ART`“ schreiben Sie bitte eine kurze Beschreibung der Art der Webseite.

In die letzte Spalte „`BEWERTUNG`“ schreiben Sie einen der folgenden Werte:

- „`++`“: Ich bin ziemlich sicher, dass dies eine nützliche Webseite ist. Ich werde sie im Laufe des Semesters sehr wahrscheinlich nochmals aufsuchen, vermutlich sogar mehrfach.
- „`+`“: Nach erstem Eindruck scheint mir dies eine nützliche Seite zu sein. Eventuell werde ich sie im Laufe des Semesters nochmal genauer lesen.
- „`o`“: Dies könnte vielleicht eine nützliche Seite sein, aber eher nicht für mich. Nach derzeitigem Stand glaube ich nicht, dass ich mir diese Seite nochmals anschauen werde.
- „`-`“: Nach meiner Einschätzung ist die Seite keine nützliche Quelle für Studierende dieser Vorlesung (auch nicht für Studierende der Fortsetzung im Sommer).
- „`--`“: Ich bin ziemlich sicher, dass diese Seite von der Link-Liste der Vorlesung gestrichen werden sollte.
- „`!`“: Der Link funktioniert nicht mehr. (Links mit dieser Art von Bewertung zählen bei den verlangten fünf Seiten aber nicht mit.)

Abgabe im Datenformat CSV:
Sie müssen die Daten im CSV-Format („comma-separated values“) abgeben. Dieses Format wird häufig benutzt, um Tabellen auszutauschen (es wäre z.B. möglich, dass die Webseite Ihrer Bank erlaubt, den Kontoauszug als CSV herunterzuladen). CSV ist auch ein Export-Format von Tabellen-Kalkulationen (wie Excel). Sie sollten die Datei aber besser mit einem normalen Texteditor erstellen, den Sie auch zur Programmierung verwenden würden (z.B. Notepad++ [https://notepad-plus-plus.org/]). Mindestens sollten Sie die erzeugte Datei in einem Editor anschauen, und sich nicht darauf verlassen, dass Excel sie schon korrekt erzeugt. Es gibt bei CSV recht viele Abweichungen vom offiziellen Format nach [RFC 4180]. Z.B. werden teils auch andere Trennzeichen als Kommata für die Spalten verwendet, wir akzeptieren aber nur Kommata. Zu der Hausaufgabe gehört auch, dass Sie sich in dieses Format einarbeiten.

In einfachsten Fall gibt es eine Zeile in der Datei pro Tabellenzeile (Datensatz), und die Werte der Spalten sind durch Kommata getrennt. Man könnte die obige Tabelle also so aufschreiben:

```
12345,0,VORRECHNEN:4,,
12345,1,https://de.wikipedia.org/wiki/SQL,Lexikon-Artikel,+
12345,2,https://www.postgresql.org/,DBMS-Homepage,++
```

Wenn ein Spalteneintrag ein Komma enthält (oder einen Zeilenumbruch) funktioniert das so natürlich nicht, dann muss man den Datenwert in `"..."` einschließen. Man darf die Datenwerte immer so einschließen, auch wenn sie nicht Kommata oder Zeilenumbrüche

enthalten. Wenn ein Datenwert ein Anführungszeichen " enthält, muss es verdoppelt werden.

Es gibt noch eine optionale Kopfzeile mit den Spaltennamen, die verwenden Sie bitte nicht. Man kann dem Datei-Inhalt nicht ansehen, ob die erste Zeile als Kopfzeile zu interpretieren ist oder als erster Datensatz. Wir haben uns dagegen entschieden.

Verwenden Sie die UTF-8 Codierung für Zeichen (das ist nur relevant, wenn Ihr Text Umlaute enthält). Schreiben Sie die Tabellendaten in eine Datei „`webseiten.csv`“, und laden Sie diese in die Übungsplattform hoch.

Wir werden versuchen, alle Abgaben mit folgendem `psql`-Befehl in eine entsprechende Tabelle zu laden:

```
\COPY WEBSEITEN FROM 'webseiten.csv' CSV DELIMITER ',' ENCODING 'UTF-8'
```

Es ist damit zu rechnen, dass es Punktabzüge gibt, wenn das nicht funktioniert. Leider können Sie es mit dem Adminer nicht testen. Falls Sie PostgreSQL selbst installiert haben, können Sie eine „`CREATE TABLE`“-Anweisung für die Tabelle unter folgendem Link abrufen:

[https://users.informatik.uni-halle.de/~brass/db24/homework/webseiten.sql]

Weitere Quellen zum CSV-Datenformat:

- Der Internet-Standard „RFC 4180“ für den MIME-Typ `text/csv`:

  [https://www.rfc-editor.org/info/rfc4180]

- Englischsprachige Wikipedia:

  [https://en.wikipedia.org/wiki/Comma-separated_values]

- Deutsche Wikipedia:

  [https://de.wikipedia.org/wiki/CSV_(Dateiformat)]

- Meine Folien zur Vorlesung „Datenbank-Programmierung“, ab Folie 3-32:

  [https://users.informatik.uni-halle.de/~brass/dbp24/slides/p3_updat.pdf]