

Übungsblatt 5: „SQL-Anfragen“

Abgabe bis Montag, 5. Juni 2023, 14:00 Uhr über das Abgabesystem im Moodle

Wichtige Hinweise zu den Aufgaben und zur Abgabe

- **Datenbank:** Für alle SQL-Anfragen auf diesem Übungsblatt wird die **“TweetDB”** verwendet. Um Ihre SQL-Anfragen zu testen und die Anfrageergebnisse zu ermitteln, müssen Sie die Datenbank in PostgreSQL importieren. Informationen zur Installation von PostgreSQL und zum Import der “TweetDB”-Datenbank finden Sie auf der Moodle-Webseite der Vorlesung. Dort finden Sie auch eine Beschreibung der Datenbank-Tabellen.
- **Abgabe:** Zu jeder Anfrage ist der SQL-Ausdruck **und das Ergebnis der Anfrage** anzugeben; bei korrekter Anfrage aber fehlenden Ergebnissen gibt es nur die Hälfte der Punkte. Für die Angabe der Ergebnisse gilt: Hat das Ergebnis mehr als 15 Tupel, ist nur die Anzahl der Tupel anzugeben. Hat das Ergebnis 15 oder weniger Tupel, so sind alle Ergebnistupel explizit anzugeben. Nicht ausführbare Anfragen (Syntaxfehler) werden mit 0 Punkten bewertet. Für semantische Fehler oder falsche Ergebnisse gibt es Abzüge. Kopieren Sie die SQL-Ausdrücke und die Ergebnisse am besten direkt aus der Konsole bzw. aus pgAdmin in Ihr Dokument, um Syntaxfehler zu vermeiden.
- **Code-Abgabe:** Bitte geben Sie zu allen Aufgaben, in denen Sie SQL-Anfragen formulieren müssen (mit ☒ markiert), immer **zusätzlich eine einfache Textdatei** (kein pdf) ab. Diese soll neben Ihren Lösungen in Ihrer abgegebenen Archivdatei existieren, ersetzt aber nicht die sonst abgegebene Lösung. Die Textdatei dient den TutorInnen ausschließlich zur Unterstützung bei der Korrektur komplexer Anfragen und soll diesen erlauben, den Code einfacher zu kopieren.

Aufgabe 5-1 Grundlegende Anfragen

2 + 2 + 3 + 4 + 4 = 15 Punkte



Formulieren Sie folgenden Anfragen als jeweils genau ein SQL **select**-Statement. **Diese Anfragen sind ohne die Verwendung von Aggregat-Funktionen, Gruppierungen und der limit-/offset-/fetch-Klausel zu formulieren.**

1. Geben Sie die Namen (`real_name`) der Accounts der Kategorie „lobby“ aus, die vor dem 30.6.2009 (UTC) auf Twitter angelegt wurden. Die Ausgabe ist dabei aufsteigend nach dem Datum (`created_at`) zu sortieren, das mit ausgegeben werden soll.
2. Gesucht sind die Namen (`twitter_name`) der Twitter Accounts, die einen Tweet gepostet haben, der zwischen 22.000 und 25.000 Likes bekommen hat. Neben dem Namen ist jeweils auch die Anzahl der Likes auszugeben.
3. Welche verschiedenen Hashtags hat der Nutzer mit dem Real Name 'LobbyControl' in dem Zeitraum 1.1.2023 bis 15.1.2023 verwendet? Die Hashtags bzw. deren Text sollen alphabetisch sortiert ausgegeben werden.
4. Gibt es Hashtags, die mindestens zwei Mal in einem Tweet verwendet wurden? Geben Sie die ID und den Text dieser Hashtags aus.

5. Von allen Twitter Accounts, die vor dem 1.1.2010 angelegt wurden, geben Sie den Namen (`real_name`) des Accounts aus, der die meisten Follower (basierend auf `follower_count`) hat. Die Anzahl der Follower soll mit ausgegeben werden. Achtung: es darf hierzu keine Aggregatfunktion verwendet werden.

Aufgabe 5-2 String-Funktionen

2 + 2 + 2 = 6 Punkte



Formulieren Sie folgende Anfragen in SQL unter Verwendung geeigneter String-Funktionen. Eine Beschreibung der Funktionen finden Sie in der PostgreSQL Dokumentation zu *String-Funktionen und Operatoren* (<https://www.postgresql.org/docs/current/functions-string.html>).

1. Ausgegeben werden sollen die Texte zu Tweets, in denen das Wort 'OpenAI' auftaucht (Groß-/Kleinschreibung spielt keine Rolle) und die mindestens 20 Mal retweetet wurden.
2. Was ist der Text zum längsten Named Entity (meisten Buchstaben)? Auszugeben ist der Text des Named Entity und dessen Länge.
3. Ausgegeben sind die Texte zu allen Named Entities, die mindestens 4 Zeichen haben und deren Text ein Palindrom bildet, also vorwärts und rückwärts gleich geschrieben wird.

Aufgabe 5-3 Exists-Operatoren

2 + 3 + 4 = 9 Punkte



Formulieren Sie folgende Anfragen in SQL unter Verwendung geeigneter `[not] exists`-Operatoren. In jeder Anfrage muss mindestens einer dieser Operatoren verwendet werden.

1. Geben Sie alle Informationen zu Named Entities aus, die in keinem Tweet verwendet wurden.
2. Geben Sie die Texte der Tweets aus, in denen sowohl der Hashtag 'klima' als auch das Named Entity 'Berlin' auftauchen.
3. Geben Sie die Namen der Nutzer aus (`real_name` und `twitter_name`), die einen Tweet nach dem 15.2.2023 gepostet haben, der mindestens 70 Tweets in seiner Conversation hat.

Aufgabe 5-4 Aggregat-Funktionen und Gruppierung

2 + 3 + 4 + 4 = 13 Punkte



Formulieren Sie folgende Anfragen in SQL unter Verwendung geeigneter Aggregat-Funktionen und Gruppierungen. **Die Anfragen sind ohne Verwendung der `limit`-/`offset`-/`fetch`-Klausel-Klausel zu formulieren.**

1. Für jedes Named Entity soll die Anzahl der Postings des Named Entity in Tweets bestimmt werden. Ausgegeben sind die IDs der Named Entities, deren Text sowie die Anzahl der Tweets, in denen das Named Entity auftaucht. Hierbei braucht nicht geprüft werden, ob ein Named Entity mehrmals in einem Tweet verwendet wird (jedes Posting zählt). Die Named Entities sind absteigend nach der Zahl der Häufigkeit ihrer Postings auszugeben.
2. Für jeden Account vom Typ 'politician' soll die Anzahl der Tweets bestimmt werden, die nach dem 1.1.2022 gepostet wurde, basierend auf den Einträgen in der Tabelle TWEET. Ausgegeben sind der Real Name sowie die Anzahl der Tweets, absteigend sortiert nach der Anzahl der Tweets. Weiterhin sollen nur Accounts ausgegeben werden, die mehr als 2000 Tweets gepostet haben.
3. Geben Sie das Erstellungsdatum und den Text des Tweets aus, in dem die meisten Named Entities verwendet wurden (hierbei darf ein Named Entity in einem Tweet auch mehrmals auftauchen, und jedes Posting wird gezählt).

4. An welchem Tag nach dem 31.12.2022 wurden die meisten Tweets gepostet? Auszugeben ist der Tag und die Anzahl der Tweets an dem Tag. Es ist in Ordnung, wenn die Ausgabe des Tages den Suffix '00:00:00+02' beinhaltet. Informationen zu Datums- und Zeitfunktionen in Postgres finden Sie unter <https://www.postgresql.org/docs/current/functions-datetime.html>.

Informationen zur Abgabe. Die Aufgaben können in Gruppen bis zu **drei** Studierende bearbeitet und abgegeben werden. Bitte schreiben Sie die Namen aller Mitglieder ihrer Gruppe sowie die Nummer ihrer Übungsgruppe (1, 2 oder 3) auf das Frontblatt ihrer Abgabe! Zur Erinnerung, hier die Übungsgruppen:

Gruppe 1: Donnerstag, 14:00 - 16:00 Uhr, Stefan Hickl
Gruppe 2: Freitag, 9:00 - 11:00 Uhr, Khadija Abdul-Latif
Gruppe 3: Freitag, 14:00 - 16:00 Uhr, Nicolai Smolkin

Reichen Sie Ihre Abgabe lediglich über die bereitgestellte Upload-Funktion in Moodle ein. Für die digitale Abgabe bieten sich elektronische Formate an. Wie es im wissenschaftlichen Rahmen üblich ist, empfehlen wir Ihnen hier die Benutzung von \LaTeX . Scans von handgeschriebenen Lösungen sind ebenfalls in Ordnung. Verwenden Sie idealerweise einen Scanner (achten Sie auf korrekte Einstellungen für DPI) oder eine Kamera mit hoher Bildqualität und hohem Kontrast. Schreiben Sie klar und deutlich und verwenden Sie keinen Bleistift.