Datenbanksysteme

Projektdokumentation

Freie Universität Berlin
Fachbereich Mathematik und Informatik, Institut für Informatik
Serkan Baris, Davit Yuzbashian, Thushan Satkunanathan
Sommersemester 2017

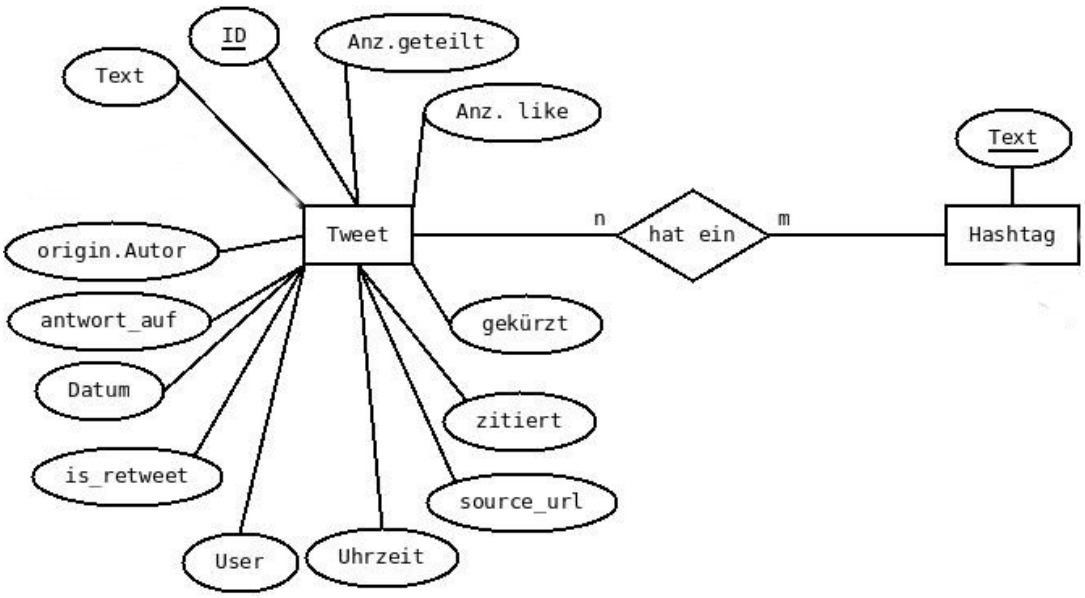
1.Iteration

Aufgabe 1)
Das Team: Serkan Baris, Davit Yuzbashian, Thushan Satkunanathan
Wir sind Informatik Studenten an der Freie Universität Berlin und möchten in diesem Projekt eine Web-Anwendung erstellen, womit wir Abfragen, Visualisierungen unserer Abfragen oder auch unserer Datenbank erstellen können.

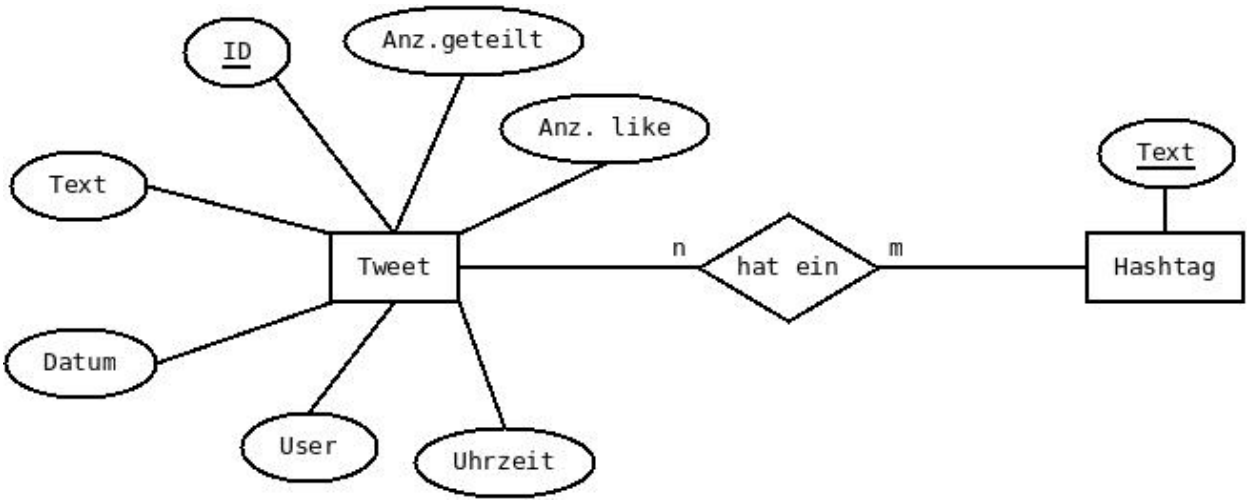
Aufgabe 2)
In der ersten Spalte steht der Benutzername von demjenigen, der diesen Tweet erstellt oder geteilt hat(”realDonaldTrump” oder ”HillaryClinton”), die zweite Spalte beinhaltet den Tweet. Gegebenfalls existiert auch ein Link. Die Spalte C besteht aus booleschen Werten, die bei True angeben ob dieser Tweet von jemand anderem stammt, wobei dann auch in Spalte D der Autor von dem orginalen Tweet ist, falls False dann existiert auch kein Eintrag in der Spalte D, denn die Spalte A gibt den Autor des Tweets an. Der Zeit-

punkt des Tweets ist in Spalte E abzulesen. Falls der Tweet eine Antwort auf einen anderen Tweet ist, dann steht der Empfänger der Antwort in Spalte F. In Spalte G können wir ablesen, ob der Tweet ein Zitat ist. Spalte H zeigt an, wie oft der Tweet geteilt wurde und Spalte I zeigt an, wie oft der Tweet geliked wurde. In der Spalte J können wir durch die URL ablesen mit was für einem Endgerät dieser Tweet erstellt wurde. Falls der Tweet gekürzt wurde, kann man dies durch einem True in Spalte K erkennen.

Aufgabe 3)



Bei diesem ER Modell bercksichtigen wir die Spalten aus der CSV Datei ebenfalls als Attribute, auch wenn sie irrelevant fr die Abfragen sind. Das folgende ER-Modell bezieht sich nur auf die Abfragen. Das Relationale Modell haben wir zu dem folgendem ER Diagramm erstellt.



Aufgabe 4)
Tweet(ID, Anz.geteilt, Anz.like, Text, Datum, User, Uhrzeit)
Hashtag(Text)
hat ein(Tweet.ID, Hashtag.Text)
Bei Tweet haben wir als Schlsselattribut ID genommen, da dies die effizienteste Methode ist. Ein Hashtag ist eindeutig, deswegen kann man als Schlsselattribut den Hashtag selbst nehmen. Bei der Relation "hat ein" haben wir als Schlsselattribut Tweet.ID und Hashtag.Text genommen, denn ein Hashtag kann in mehreren Tweets auftreten und ein Tweet kann mehrere Hashtags haben.

Aufgabe 5)
Mit dem Befehl "sudo -u postgres createdb Election" können wir eine Datenbank mit dem Namen Election und dem default Benuter postgres erstellen.

2.Iteration

In der 2. Iteration war die Hauptaufgabe die Daten aus der CSV bzw Excel Datei einzulesen und in die von uns erstellten Datenbank einzufügen. Dabei hatten wir zuerst unser Datenbankschema mithilfe der Create Statements erstellt.

Zur Datenbereinigung Nutzen wir die Programmiersprache Java, da wir bereits Erfahrung damit erarbeitet haben und Eclipse als Entwicklungsumgebung, viel Arbeit in Sachen recherchieren abnimmt, weil man für alle mglichen Funktionen eine Beschreibung geliefert bekommt. Zusätzlich war es notwendig Apache Poi herunterzuladen, da dies die notwendige Java-Programmbibliothek ist, um Daten im Dateiformat wie z.B. Word und Excel zu lesen oder zu schreiben. Wir entschieden uns fr die Excel Datei anstatt der CSV, da die CSV beim einlesen unbekannte Zeichen (Schwarze Raute mit Fragezeichen) beinhaltete, dieses Problem wollten wir beheben. Das einlesen der Excel Datei erwies sich anfangs als sehr schwierig, da uns nicht ganz schlssig war, wie genau die Daten eingelesen werden, jedoch erwies sie sich als weitaus schnere Variante.

Als nächstes mussten wir die Daten der Excel Tabelle in unsere Datenbank importieren. Auch hierzu benutzen wir Java als Programmiersprache. Beim Einfügen der Daten mussten wir unser ER-Modell aus der 1.Iteration wiederaufgreifen. Dabei fiel uns auf dass wir z.B. Das Datum, welches im format (YYYY-MM-DD T Uhrzeit) war trennen mussten, da wir Uhrzeit und Datum als separate Attribute betrachten. Des Weiteren werden leere Zellen von der Excel Datei nicht als leere Felder bergeben, sondern direkt ignoriert, welches zu Problemen fhrte. Aber auch die Semikolons mussten wir aus den Tweets entfernen, da es sonst Problem in PostgreSQL gbe, wenn die Daten eingefgt werden sollen, da die Semikolons dort als Anfang bzw Ende eines Strings erkannt werden.

Link zum Code: `https://github.com/sekores/Election/blob/master/Database.java`

Zuletzt haben wir noch einen Webserver aufgesetzt, und zwar denselben, den wir bereits in dem 1. bungsblatt erstellt haben. Dieser wre Apache 2. Wir entschieden uns dafr aus folgenden Grnden. War das fr uns am naheliegendsten, da wir whrend der Bearbeitung dieser Iteration, Probleme mit dem Zeitmanagement hatten und die Vorarbeit aus dem bungsblatt uns Zeit verschafft hat an der Programmieraufgabe zu sitzen.

Link zu dem Repository: `https://github.com/sekores/Election`