Datenbanksysteme
Projekt Iteration 1: Modellierung

Adrian Gruszczynski
Florian Brinkmeyer
Pit Ronk
Remi Toudic
Tutor: Christian Hofmann
Tutorium: Donnerstag 12-14

May 11, 2017

Aufgabe 1: Projektdokumentation

Das Team in der alphabetischen Reihenfolge:

- Adrian Gruszczynski (Informatik Bachelor)
- Florian Brinkmeyer (Informatik Lehramt)
- Pit Ronk (Mathematik Bachelor)
- Remi Toudic (Mathematik Master)

Projektziel

Das Ziel des Projektes, ist die Entstehung einer Web Anwendung. Sie soll dabei Daten aus dem Datensatz *american-election* in geeigneter Form, visuell darstellen sowie das Abfragen von Informationen ermöglichen.

Das sekundäre Ziel ist dabei, das Auseinandersetzen mit relationale Datenbanken und deren Anwendung.

Des Weiteren ist der Weg ebenso ein Ziel, welcher sich aus folgenden Punkten zusammen setzt.

Phase 1: Hierbei liegt der Fokus auf der Festlegung der zu implementierenden Features sowie das Entwerfen des Konzeptes für das Datenmodell.

Phase 2: Erstellung des Datenbankschemas sowie die Aufbereitung und das Pflegen der Daten. Des Weiteren folgt Einbindung in den Webserver.

Phase 3: Bestimmung der Datenabhängigkeit sowie dessen visuelle Darstellung.

Aufgabe 2: Explorative Datenanalyse

Die einfache Analyse

Die einfache Analyse

Der Datensatz *american-election-tweets.xlsx* besteht aus 6127 Einträgen und erfasst Tweets von Hillary Clinton und Donald Trump, sowie die Metadaten aus der Zeiten des Präsidentenwahl 2016 in Amerika.

Der Datensatz wird durch folgende Attribute charakterisiert:

- *handle* : beschreibt den username des Tweet Authors
- *text* : speichert den textlichen Inhalt des Tweets
- *is_reetwet* : sagt aus ob der jeweilige Tweet retweeted wurde. Kann nur boolean Werte annehmen
- *original_author* : beschreibt den ursprünglichen Author des Tweets in falle eines retweets
- *time:* speichert den Zeitstempel jedes Tweets
- *in_reply_to_screen_name* : falls der Tweet ein reply ist speichert der Parameter den ursprünglichen Author
- *is_quote_status* : besagt ob der Tweet ein Zitat eines anderes Tweets ist. Kann nur boolean Werte annehmen

- *retweet_count*: speichert wie viel Mal der jeweilige Tweet repostet/retweeted wurde
- *favorite_count*: speichert wie viel Mal der jeweilige Tweet geliked wurde
- *source_url*: beschreibt die Quelle des Tweets

Die explorative Analyse

Der Datensatz beinhaltet Informationen über Tweets im Zeit des 01.05. bis 28.09.2016. Die Thematik behandelt die Präsidentenwahl in Amerika. Es gibt insgesamt 6126 Einträge von denen ca. 50.3% von Hillary Clinton und 49.7% von Donald Trump erstellt wurden.

Die Tweets von Hillary Clinton wurden insgesamt 9346511 Mal retweeted und 21026005 Mal als Favorit markiert. Die Tweets von Donald Trump wurden 17668486 Male retweeted und 50778572 Male als Favorit gekennzeichnet.

Obwohl es im dem Datensatz etwas mehr Tweets von Hillary Clinton gibt, wurden die Tweets von Donald Trump fast doppelt oft geliked und retweeted. Die Anzahl der Retweets von Donald Trump macht somit 65% der gesamten Retweets und 70% der ganzen likes aus dem gegebenen Datensatz aus.

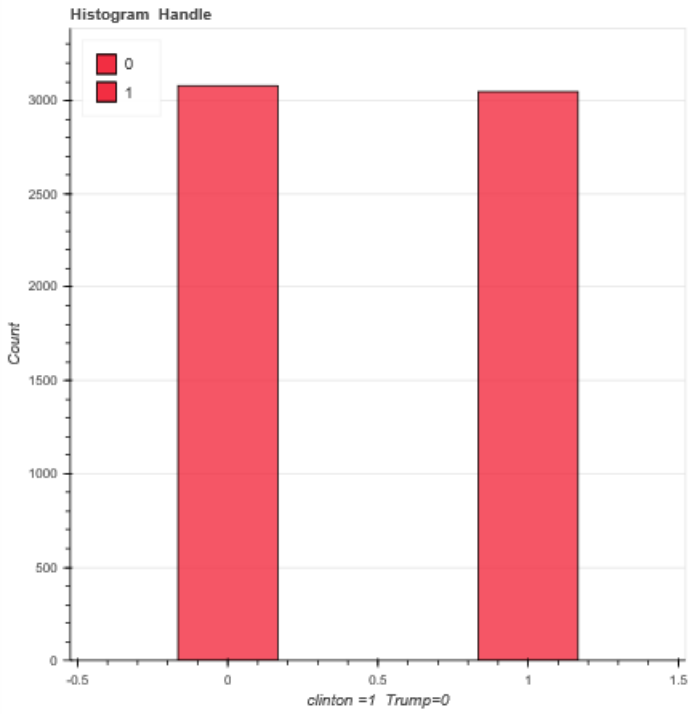
Ein durchschnittlicher Tweet von Trump wurde 16.665 Male retweeted und 5.798 Male als favorite markiert. Ein durchschnittlicher Tweet von Clinton wurde 3.035 Male retweeted und 6.828 Male geliked.

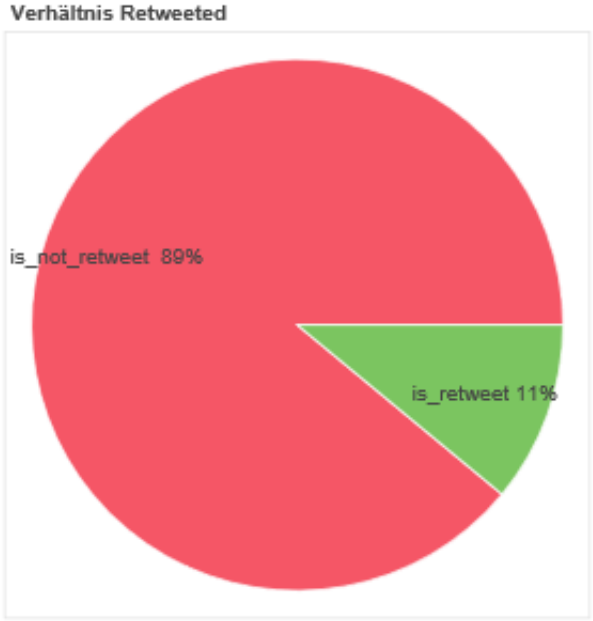
Zusammenhänge der Daten

Die Korrelationsmatrix der wichtigsten Parameter:

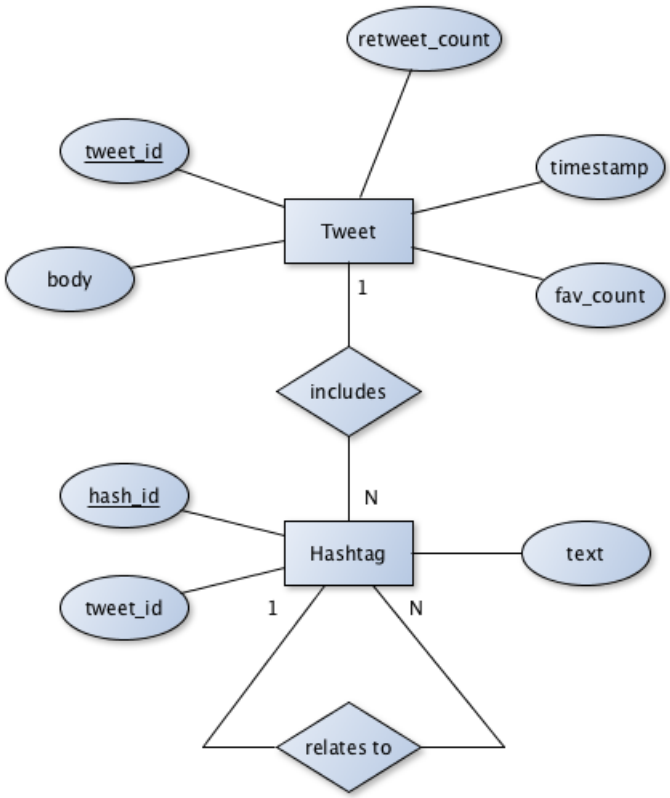
	is_retweet	s_quote_statu	retweet_count	favorite_count
is_retweet	1	0.031244	-0.060543	-0.120580
is_quote_status	0.031244	1	0.063866	0.024553
retweet_count	-0.060543	0.063866	1	0.927322
favorite_count	-0.120580	0.024553	0.927322	1

Die Anzahl der jeweiligen Tweets von Clinton und Trump werden graphisch wie folgt dargestellt:





Aufgabe 3: ER-Modellierung



Das ER-Modell besteht aus 2 Entitätstypen, die die Datensätze aus jeweils gegebenen Attributen enthalten. Der Entitätstyp 'Tweet' verfügt über folgende Attribute:

- *tweet_id* ist ein Primärschlüssel der die Entitäten innerhalb des Entitätstyps eindeutig identifitzieren lässt.
- *body*: speichert den Inhalt jedes Tweets
- *fav_count*: speichert die Anzahl der Likes für jede Instanz und lässt somit die Popularität eines Tweets abschätzen
- *retweet_count*: speichert die Anzahl der Retweets für jede Instanz, kann als Kriterium sowohl für Popularität als auch Wichtigkeit eines Tweets benutzt werden

- *timestamp*: speichert den timestamp (Zeitstempel) von jedem Tweet und kann im späteren Stadium dazu dienen, die Entwicklung von Hashtags über die Zeit zu analysieren

Der Entitätstyp Hashtag kapselt folgende Attributen:

- *hash_id*: ist ein Primärschlüssel der die Entitäten innerhalb des Entitätstyps eindeutig identifizieren lässt
- *tweet_id*: ist ein Fremdschlüssel der jeder Hashtag Instanz den dazugehörigen Tweet zuordnen lässt
- *text*: speichert den Inhalt von jedem Hashtag

Jede Tweet Instanz kann, muss aber nicht, einen oder mehrere Hashtag Instanzen beinhalten. Jeder Hashtag Instanz wird anhand des Fremdschlüssels genau eine Tweet Entität zugeordnet.

Die Relation *includes* ermöglicht dementsprechend für jeden Hashtags das dazugehörige Tweet zu zuordnen, und auch im späteren Stadium Zugriff auf die Attribute der beiden Entitäten.

Jede Hashtag Entität kann, muss aber nicht, mit einem oder mehreren Hashtags Entitäten in Relation stehen. Wenn zwei oder mehr Hashtags die selbe *tweet_id* zugeordnet wird bedeutet das, dass sie in Relation zueinander stehen, und tauchen zusammen in einem Tweet auf.

Die Relation *relates_to* erlaubt die Bestimmung des paarweisen Auftretens mehrerer Hashtags in einem Tweet.

Das obere ER-Model sowohl als auch das untere relationale Model, beinhalten nur ausgewählte Attribute aus dem gegebenen Datensatz. Da die restlichen Attribute wie *is_retweet*, *is_quote_status* usw. für die Anwendung unrelevant zu sein scheinen, haben wir uns entschieden diese vorerst zu ignorieren.

Aufgabe 4: Relationales Modell

$TWEET\{[tweet\_id : INT\ PRIMARY\ KEY, fav\_count : INT, timestamp : DATETIME(YYYY - MM - DDThh : mm : ss), retweet\_count : INT, body : TEXT(200)]\}$

$HASHTAG\{[hash\_id : INT\ PRIMARY\ KEY, text : CHAR(30), tweet\_id : INT\ FOREIGN\ KEY]\}$

Die Relation TWEET beinhaltet die wichtigsten Attribute aus dem gegebenen Datensatz. Anhand dessen, kann die Wichtigkeit und die Polarität sowie Datum als auch Inhalt des Tweets bestimmt werden. Jedes Tupel wird durch eine eindeutige ID Nummer bezeichnet die für jede Instanz als Primary Key verwendet wird.

Die Relation HASHTAG besteht aus einer *hash_id* die als Primary Key zur Identifizierung der Instanzen dient sowie anderen Attributen die den Inhalt und dazugehörige *tweet_id* beschriften.

Aufgabe 5: Datenbank erstellen

```
1  hidden@hidden-VirtualBox:/var/www/html$ sudo -i -u
       postgres
2  postgres@hidden-VirtualBox:~$ ls
3  9.6
4  postgres@hidden-VirtualBox:~$ psql
5  psql (9.6.2)
6  Type "help" for help.
7  postgres=# create database election;
8  CREATE DATABASE
9  postgres=# \list
10                                   List of databases
11     Name    |  Owner   | Encoding |   Collate   |
          Ctype    |   Access privileges
12 -----------+----------+----------+-------------+-------------+-----------
```

```
13 | election  | postgres | UTF8     | en_US.UTF-8 | en_US.
     UTF-8 |
14 | postgres  | postgres | UTF8     | en_US.UTF-8 | en_US.
     UTF-8 |
15 | template0 | postgres | UTF8     | en_US.UTF-8 | en_US.
     UTF-8 | =c/postgres          +
16 |           |          |              |
                          | postgres=CTc/postgres
17 | template1 | postgres | UTF8     | en_US.UTF-8 | en_US.
     UTF-8 | =c/postgres          +
18 |           |          |              |
                          | postgres=CTc/postgres
19 | (4 rows)
```