Universität Heidelberg
Arbeitsgruppe Datenbanksysteme
Prof. Dr. Michael Gertz
John Ziegler, Nicolas Reuter

19. Mai 2023
Datenbanken
Sommersemester 2023

Beispieldatenbank „TwitterDB"

In den weiteren Übungen werden wir häufig eine Datenbank verwenden, an die Anfragen in SQL zu formulieren sind. Diese Datenbank sollten Sie sobald wie möglich unter PostgreSQL installieren, so dass Sie Ihre SQL-Anfragen direkt testen können. Informationen zur Installation der Datenbank stehen auf den Kurs-Webseiten in Moodle zur Verfügung. Achtung: Die unkomprimierte Datei mit den Daten zur Datenbank hat eine Größe von ungefähr 100 MB!

Bei der Datenbank handelt es sich um einen Auszug aus den Daten, die im Rahmen des EPINetz-Projekts (`https://epinetz.de`) gesammelt und analysiert werden. Der Datenauszug umfasst Twitter-Daten für die Monate Januar, Februar und März 2023, also einen Zeitraum Anfang diesen Jahres. Die Daten beinhalten Informationen zu Tweets, Hashtags und Named Entities (Personen, Organisationen und Orte) sowie zu Twitter Accounts.

Nachfolgend sind die **sieben** Tabellen dieser Datenbank mit dem Namen TwitterDB mit Beispielausprägungen (d.h. ein paar Beispieltupel) angegeben, und jede Tabelle mit ihren Attributen wird kurz erläutert. Die Schlüsselattribute sind in den Tabellen unterstrichen.

Twitter User. Informationen zu Twitter User bzw. Accounts werden in der Tabelle `TWITTER_USER` verwaltet. Jeder User hat eine eindeutige `id` (vom Typ `bigint`), eine Anzahl an Follower (`follower_count`) und eine Anzahl an Tweets (`tweet_count`), die bisher von diesem Account gepostet wurden. In der Datenbank wird zwischen Accounts von politischen Organisationen/Politiker:innen und Accounts von Lobbyist:innen unterschieden. Hierzu dient das Attribut `typ`, das nur die Werte `politician` und `lobby` annehmen kann.

Zu einem Account wird auch vermerkt, wann dieser auf der Twitter-Plattform angelegt wurde (`created_at`). Diese Werte werden basierend auf dem Datentyp `timestamp with time zone` (UTC) gespeichert, siehe auch `https://www.postgresql.org/docs/current/datatype-datetime.html`. Zu einem Account werden schliesslich noch der Twitter-Name (wie er z.B. in @mentions verwendet wird) und der echte Name (`real_name`) basierend auf den Metadaten zu dem Account verwaltet.

TWITTER_USER						
id	follower_count	tweet_count	typ	created_at	twitter_name	real_name
5876652	761419	6099	lobby	2007-05-08 23:10:26+02	saschalobo	Sascha Lobo
3292982985	1060395	10212	politician	2015-05-21 16:01:21+02	Karl_Lauterbach	Prof. Karl Lauterbach
373872419	3628	5582	politician	2011-09-15 11:51:07+02	GrueneLtRLP	GrueneLandtagRLP

Tweets. In der Tabelle `TWEET` werden die Daten zu (fast nur deutschsprachigen) Tweets verwaltet. Ein Tweet hat eine eindeutige `id` (wieder vom Typ `bigint`), einen Retweet Count, einen Reply Count und einen Like Count. Ein Tweet ist über das Attribut `author_id` einem Twitter Account zugeordnet. Über das Attribut `created_at` wird nachgehalten wann ein Tweet gepostet wurde. Der Text eines Tweets findet sich als Attributwert zu `txt`.

Die Tabelle `TWEET` beinhaltet „initiale" Tweets, Replies, Quotes und Retweets, und jeder Tweet hat eine Conversation ID. Für einen „initialen" Tweet ist es die ID des Tweets selbst. Man kann diesen Wert als ID eines Thread verstehen, so dass es mehrere Tweets mit der gleichen Conversation ID geben kann, die dann zu diesem Thread gehören, wie in der nachfolgenden Tabelle weiter erläutert. Beispiele zu Einträgen dieser Tabelle finden Sie auf der letzten Seite.

Conversation. Wie oben angemerkt hat jeder Tweet eine Conversation ID. Quotes, Comments und Retweets zu einem initialen Tweet haben alle die gleiche Conversation ID wie der originale Tweet (Attribut `conversation_id` in der Tabelle `TWEET`). In der Tabelle `CONVERSATION` wird für jeden Tweet mit einer Identifikation `id` nachgehalten, welche Quotes, Comments und Retweets (welches ja auch Tweets sind) gepostet wurden. Die IDs dieser Tweets werden in Form eines Arrays `tweets` mit `bigint`-Werten dargestellt.

CONVERSATION

id	tweets
1641508538999603200	{1641861332558381056,1641508538999603200}
1641921283431563265	{1641921283431563265}
1641806378590150656	{1641857922618515458,1641856875657322516,1641806378590150656}
1641865062301196303	{1641915061873549313,1641914628228759553,1641865062301196303}
1641844591618818066	{1641844591618818066 }

Hashtags. Für viele Analysen sind die in den Tweets verwendeten Hashtags von Interesse, da sie eine gewisse Thematik widerspiegeln. In der Tabelle `HASHTAG` wird zu jedem Hashtag eine eindeutige ID (`id`) und der Text (`txt`) des Hashtags selber abgespeichert, lower-cased und ohne das Präfix '#'.

Verwendung von Hashtags. In der Tabelle `HASHTAG_POSTING` wird verwaltet, welcher Hashtag (`hashtag_id`) in welchem Tweet (`tweet_id`) an welcher Stelle (`pos_start`, entspricht dem offset) im Text auftritt. Zu beachten ist hier die Zusammensetzung des Primärschlüssels der Tabelle.

HASHTAG

id	txt
19	silvester2023
7432	ftp
6474	zugspitze
9230	bundesratspräsident
9491	energiesicherung

HASHTAG_POSTING

tweet_id	hashtag_id	pos_start
1609459544962338818	144	50
1609463158711947264	148	4
1609464785946509313	158	95
1609466393761832967	167	74
1609466393761832967	170	275

Named Entities. Bei vielen Analysen sind auch die in den Texten vorkommenden Named Entities von Interesse. Dies sind typischerweise Namen von Personen, Organisation, Unternehmen und Orten. Diese werden in der Tabelle `NAMED_ENTITY` verwaltet, wobei neben dem Text (`txt`) die eindeutige ID (`id`) zu einem Named Entity abgespeichert wird.

Vorkommen von Named Entities. Analog wie für Hashtags wird auch das Vorkommen von Named Entities in Tweets bzw. deren Texte verwaltet. Dazu dient die Tabelle `NAMED_ENTITY_POSTING`.

NAMED_ENTITY

id	txt
1391	Python
2502	Neukölln
3605	Stanford
4992	ML
299	Omicron

NAMED_ENTITY_POSTING

tweet_id	named_entity_id	pos_start
1609499123413127170	188	155
1609557999194083328	28	69
1609536177761128454	223	205
1609342276563525638	31	18
1609535386916732928	307	153

Anmerkung: Um ein besseres Verständnis für den Aufbau und die Inhalte der Tabellen zu bekommen, sollten Sie sich neben dem Schema die Daten mithilfe von einfachen SQL-Anfragen an die verschiedenen Tabellen genauer ansehen. Insbesondere sollten Sie sich Funktionen in SQL ansehen, die auf den Datentypen `timestamp with time zone` und `array` basieren (siehe `https://www.postgresql.org/docs/current/arrays.html`).

TWEET							
id	conversation_id	author_id	retweet_count	reply_count	like_count	created_at	txt
1641933672008884226	1641933672008884226	88630242	118	381	519	2023-03-31 22:41:07+00	Angesichts der real existierenden SPD und ihres Kanzlers ist es umso beeindruckender, was Robert #Habeck und @GrueneBundestag bei #Klimaschutz & #Energiewende alles hinbekommen haben: Ausbau #Erneuerbare, Aus für Atom, Kohle, Verbrenner, Gas- & Ölheizun gen https://t.co/l9we7ycu7H
1641912941195689984	1641912941195689984	433766266	0	2	5	2023-03-31 21:18:45+00	Bei Energydrinks geht die Hauptwirkung von Koffein aus, auch die Zuckermenge spielt eine Rolle. Für die Wirkung von Taurin ist di e Studienlage uneinheitlich. https://t.co/wbUmYH0iWN
1641911148986486785	1641911148986486785	399257453	0	0	1	2023-03-31 21:11:37+00	Gut gegen den #Tabellenführer gespielt, aber Zuwenig zwingende Torchancen.. jetzt Focus auf das #dfb-#pokal #Viertelfinale gegen den #vfb #stuttgart! (U+26BD)(U+26BD)(U+26BD) #fcn #fcbcvd #wirsindderclub #glubb (U+1F534)(U+26AB) https://t.co/f3PpMgziCI
1641912090234322944	1641912090234322944	2472093596	1	0	7	2023-03-31 21:15:22+00	Eine Ära geht zu Ende. Nach 14 Jahren Amtszeit haben wir heute den Rektor der @unirostock @wolfgangschare1 verabschiedet. Brücken bauer, Netzwerker und Impulsgeber - so hat er die Universität Rostock und die Hanse- und Universitätsstadt @HROrathaus entscheidend geprägt. Danke (U+1F64F) https://t.co/PeyxKy6KmQ
1641896822376484879	1641896822376484879	1425265488	8	21	110	2023-03-31 20:14:42+00	Kälte. https://t.co/UvGCi6rIrr

Tabelle 1: Beispiel-Tweets mit ihren Attributwerten; alle Tweets sind „initiale" Tweets, d.h. die Conversation ID entspricht hier der ID des Tweets selbst. Würde es sich z.B. um ein Reply handeln, wäre die Conversation ID dieses Replies (Tweet) verschieden von der ID des Tweets.