INDIVIDUAL ASSIGNMENT 1

BT2102 Data Management and Visualisation · Semester 2 · AY2017/18

OVERVIEW

This assignment provides a scenario where data needs to be captured for subsequent analysis. Our objectives are to practice your ability to develop a data model from such a scenario - from creating it conceptually, to designing its structure in a (relational) database, and implementing in it MySQL. Finally, the scenario poses a set of data requests where you will practice developing queries from relational databases using SQL.

REQUIREMENTS

For the scenario provided below, create:

1. An ER diagram modelling the data requirements of the scenario (with clear indicators of primary and foreign keys, and other relational or data constraints).
2. The list of normalised tables and fields (with data types), from this ER diagram, that can be implemented into a relational DBMS (like MySQL).
3. The MySQL instructions you would use to create each table of the database.
4. (Optional) A free-text list of any non-obvious assumptions you may be making, or limitations you would like to acknowledge of your conceptual or logical model, or other information you think is necessary for the evaluation of your model.
5. Your answers to the SQL queries posed with each scenario.

Package these requirements into a submission file (pdf), with your name and student number, and upload these to the Submission folder in IVLE. The deadline for the submission is the night of **Friday 27 April 2018.**

QUESTION 1: TEXT MINING TWITTER STREAMS

Twitter's Real-Time Streaming API allows anyone to collect tweets in real-time directly from the service. Once an API connection has been established with Twitter, tweets can be continuously delivered and processed (in batches) accordingly.

The NUS Social Media Analytics Team (SMAT) uses the API to capture tweets related to NUS as they are posted on Twitter. Once captured from Twitter, SMAT (a) processes the tweets and (b) stores them for subsequent (sentiment and opinion) analysis. The processing involves extracting relevant details from the tweet stream sent from Twitter, categorising the tweets based on how they are relevant to NUS, and tagging them with a sentiment category. Storing them involves saving these details into an RDBMS.

Receiving Tweets from Twitter

A tweet is essentially a string of text at most 280 characters in length, created at a specific date and time, at a particular location (if included with the tweet, location is captured as latitude and longitude) by a particular user, identified by a unique username and userID.

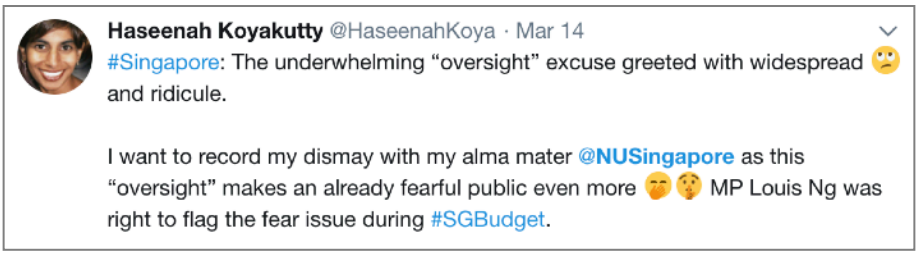

Fig. 1: An example of a tweet

Besides the text string, a tweet may also contain four other entities, specifically:

A. User hashtags - plain text strings that begin with a "#" sign
B. External URLs - plain text strings that link to websites and other web resources, outside of Twitter
C. Twitter media URLs - plain text strings that link to videos/images uploaded directly to Twitter

A tweet may have several of these three entities (eg. multiple links, multiple hashtags, and/or multiple Twitter images or videos).

Each tweet also contains some basic information about it, including a creation date, time and 2 digit code for time zone, a posting date, time and time zone code, a very large ID number that uniquely identifies the tweet, the location from where the text was posted (country, longitude, latitude), and some of the user's profile data (an ID, username, full name, url, profile description, number of followers, number of friends, number of tweets, and profile image).

A tweet may also be a retweet. A retweet is technically 2 tweets from 2 users, but they have the same text body and entities. The "original" only differs from the "re" tweet in the user that sent it and a timestamp. When retweets are found in the tweet stream from Twitter, only the full tweet of the original needs to be stored. The "re" tweets only need store the retweeting user and timestamp.

A tweet might also include "mentions", which are references (from a single *origin* user) to other Twitter users (a set of *target* users). Such mentions often connect a series of tweets into a conversation.

Finally, information about the tweet stream itself accompanies the stream, including an ID number for the batch of tweets, and the date and time of this batch. All other data in the tweet stream is ignored or dropped.

Processing a tweet

When the tweet is received, SMAT does some simple processing and decision making on each tweet. In particular, it categorises the tweet based on its sentiment - into only one of 3 categories (positive, neutral, negative). It also tags the tweets based on how they are related to NUS. There are currently a list of 24 campus-tags (one of each faculty, "undergrad", "postgrad", "education", "outreach", "events", etc), and a tweet may have several campus-tags. Furthermore, the list of campus-tags changes (grows/shrinks) with time.

Queries

After storing all this data in an RDBMS, SMAT frequently queries parts of this database for their social media dashboards or analyses. Provide the SQL queries that would extract the relevant information for the needs specified below:

A. For every country available in our data, how many tweets originated from there?
B. What are the 10 most common user hashtags used in our collected tweets?
C. Who is the most influential Twitter user in our stream (you are free to assume a measure for influence)?
D. How many tweets are there for each sentiment?
E. Which faculty has the most tweets with a positive sentiment, and which has the least?
F. Open Day was held from Fri Mar 9 to Sun Mar 11, 2018, during which time, a campaign for Twitter users encouraged posting using the hashtag "#NUSOpenDay18". For such tweets during this period, what were the most common hashtags (other than NUSOpenDay18)? What were the top 3 retweeted tweets and who were their users? What was the breakdown of the sentiment across these tweets?