Database Systems, CSCI 4380-01
Homework # 1
Due Friday January 26, 2018 at 2:00:00 PM

Homework Statement. This homework is worth 4.5% of your total grade. If you choose to skip it, Midterm #1 will be worth 4.5% more. Remember, practice is extremely important to do well in this class. I recommend that not only you solve this homework, but also work on homeworks from past semesters. Link to those is provided in the Piazza resources page.

This homework aims to teach you how to construct complex queries using relational algebra. Please do the parts in sequence. The questions get harder and build on your knowledge of relational algebra from previous parts. Each question is equal weight.

Database Description. Suppose you are given the following database that is a slightly simplified version of Reddit data. Users create `posts` with given title and text (`ptext`) for a specific subreddit (`subname`) and are given scores.

Users post comments. For simplicity, we ignore threads and associate comment with a post (`postid`). Each comment and post has an associated user (`uname`).

Users can have multiple flairs and trophies. Subreddits may have multiple users who are moderators. Users have a reddit birthday (`reddit_bday`) that is given by a date in the form `mon-day-year`. While in reality subreddits don't have a `type`, we are including one here for querying purposes.

`posts(`<u>`postid`</u>`, title, ptext, pdatetime, subname, uname, score)`

`comments(`<u>`cid`</u>`, ctext, postid, uname, cdatetime)`

`users(`<u>`uname`</u>`, email, karma, reddit_bday, reddit_gold)`

`trophies(`<u>`uname, trophyname`</u>`)`

`flairs(`<u>`uname, flairname`</u>`)`

`subreddits(`<u>`subname`</u>`, url, nummembers, created_date, description, type)`

`moderators(`<u>`subname, uname`</u>`)`

Note: All datetime fields are formatted as `mon-day-year-time`, e.g. `01-31-2016-14:00` and date files are formatted as `mon-day-year`. You can assume that you can check if a datetime or date value `X` comes after another value `Y` by checking whether `X` $>$ `Y`.

Write the following queries using relational algebra (pay attention to the attributes required in the output!):

Question 1. The following queries only need a single SELECT (σ), followed by a PROJECT (π) and RENAMING (ρ) as necessary:

(a) Return the id and title of all posts with a score above `1,000`, posted to the subreddit named `PerfectTiming` in `2017`.

(b) Return the url of all subreddits with at least `100` members and are of type `news`.

Question 2. The following queries combine SELECT (σ), SET operations ($\cap, \cup, -$), PROJECTION (π) and RENAMING (ρ) as necessary:

(a) Find and return the username of all users who either have one or more flairs or they have at least one reddit gold and one or more trophies.

(b) Return the name of subreddits with at least 200 members created in 2017 with no moderators.

Question 3. The following queries combine SELECT (σ) statements with any number of JOINS as neede ($\bowtie$, theta or natural) (or CARTESIAN PRODUCT), followed by a PROJECT (π) and RENAMING (ρ) as necessary:

(a) Return the id of all posts posted in 2017 by a user with exactly 10 reddit gold on a subreddit of type `news`.

(b) Return the id and title of all posts on subreddit named `TIL` with at least one comment posted by the same user who originated the post and has at least one comment by a different user.

(c) Return the URL of all subreddits moderated by at least one user both with a trophy and at least two flairs.

Question 4. Freeform, you decide which combination is needed. Any relational algebra operator is fine. Remember to construct these in parts and provide comments on what each part is computing. This will make it possible for us to give partial credit.

(a) Find and return the id and title of all posts posted in a moderated (i.e. has a moderator) subreddit of type `ask` with at least one comment by a user with a trophy and a flair.

(b) Find and return the username and email of all users who are moderating a subreddit of type `ask` or have created a post in a subreddit of type `ask` in `2017` with at least `1,000` points.

(c) Find and return the title, post date time of all posts with no comments, and with a negative post score, and are posted by a user with a flair who is a moderator of the subreddit the post is on.

If you are finished with all these queries but find yourself in need of a personal challenge, try to write this query to explore the expressive power of relational algebra (no hw credit for this question): Find and return the email of users who have posted in <u>all</u> subreddits with exactly two moderators, and nowhere else. (No credit for this question, try it to challenge yourself. It will be a long query for sure, so break it down to pieces.)

SUBMISSION INSTRUCTIONS. Submit a PDF document for this homework using Gradescope. No other format and no hand written homeworks please. No late submissions will be allowed.

The gradescope for homework submissions will become available by Monday January 22 the latest. We will announce it on Piazza.