Database Systems, CSCI 4380-01 Spring 2018
Homework # 1 Answers

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(postid, title, ptext, pdatetime, subname, uname, score)
    comments(cid, ctext, postid, uname, cdatetime)
    users(uname, email, karma, reddit_bday, reddit_gold)
    trophies(uname, trophyname)
    flairs(uname, flairname)
    subreddits(subname, url, nummembers, created_date, description, type)
    moderators(subname, uname)

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!).

Sample Answer. (See example Latex code below for formatting suggestions. The & signs allow you to align multiple lines. For super long conditions, you can just write them on a separate line for better readability. Please consider TAs reading these queries and make them easy to digest.)

$$
\begin{aligned}
R1 &= \Pi_{uname}\left(\sigma_{trophyname=PostgreSQLWizard}(Trophies) \bowtie (Moderators)\right) \\
R2 &= (R1 \bowtie (\sigma_{type=database}(subreddits)) \\
R3 &= \Pi_{url}(R2 \cup (\sigma_C subreddits)
\end{aligned}
$$

where $C : type = postgreSQL$ and $(nummembers > 10$ or $type = info)$.

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

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

Answer.

$$R1 \quad = \quad \Pi_{postid,title} \left(\sigma_{score \geq 1000 \text{ and } subname=PerfectTiming \text{ and } pdatetime \geq X \text{ and } pdatetime \leq Y} (Posts) \right)$$

where $X = $ 1-1-17-00:00 and $Y = $ 12-31-17-23:59.

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

Answer.

$$R1 \quad = \quad \Pi_{url} \left(\sigma_{nummembers \geq 100 \text{ and } type=news} (subreddits) \right)$$

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

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

Answer.

$$R1 \quad = \quad \Pi_{uname}(Flairs) \cup (\Pi_{uname}(\sigma_{reddit\_gold \geq 1} Users) \cap (\Pi_{uname} Trophies))$$

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

Answer.

$$R1 \quad = \quad \Pi_{subname}(\sigma_{nummembers \geq 200 \text{ and } created\_date \geq X \text{ and } created\_date \leq Y} Subreddits) - (\Pi_{subname}(Moderators))$$

where X=01-01-2017 and Y=12-31-2017.

Question 3. The following queries combine SELECT (σ) statements with a JOIN ($\bowtie$) (or CARTE-SIAN PRODUCT), followed by a PROJECT (π) and RENAMING (ρ) as necessary:

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

Answer.

$$R1 \quad = \quad \Pi_{postid}(\sigma_{pdatetime \geq X \text{ and } pdatetime \leq Y} Posts \bowtie (\sigma_{reddit\_gold=10} Users) \bowtie (\sigma_{type=news} Subreddits)$$

where $X = $ 1-1-17-00:00 and $Y = $ 12-31-17-23:59.

3(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.

Answer.

$$
\begin{aligned}
R1 \quad &= \quad \Pi_{postid}(\sigma_{subname=TIL} Posts) \bowtie Comments \\
C2(postid2, uname2) \quad &= \quad \Pi_{postid,uname} \\
R2 \quad &= \quad \Pi_{postid,title}(R1 \bowtie_{uname<>uname2 \text{ and } postid=postid2} C2)
\end{aligned}
$$

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

Answer.

$$
\begin{aligned}
F2(U2, f2) \quad &= \quad Flairs \\
R1 \quad &= \quad \Pi_{uname}(Flairs \bowtie_{uname=u2 and f2<>flairname} F2) \cap \Pi_{uname}(Trophies) \\
R2 \quad &= \quad \Pi_{url}(Subreddits \bowtie Moderators \bowtie R1)
\end{aligned}
$$

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.

4(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 a comment by a user with a trophy and a flair.

Answer.

$$
\begin{aligned}
R1 &= Posts \bowtie (\sigma_{type=ask} Subreddits) \bowtie \Pi_{subname}(Moderators) \\
R2 &= \Pi_{postid}(Comments \bowtie Trophies \bowtie Flairs) \\
R3 &= \Pi_{postid,title}(R1 \bowtie R2)
\end{aligned}
$$

4(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.

Answer.

$$
\begin{aligned}
R1 &= \sigma_{type=ask} Subreddits \\
R2 &= \Pi_{uname,email}(Users \bowtie R1 \bowtie Moderators) \\
R3 &= \Pi_{uname,email}((\sigma_{score \geq 1000 \text{ and } pdatetime \geq X \text{ and } pdatetime \leq Y} Posts) \bowtie Users \bowtie R1) \\
R4 &= R2 \cup R3
\end{aligned}
$$

where $X = $ 1-1-17-00:00 and $Y = $ 12-31-17-23:59.

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

Answer.

$$
\begin{aligned}
R1 &= \Pi_{postid}((\sigma_{score<0} Posts) \bowtie Users \bowtie Flairs \bowtie Moderators) \\
R2 &= \Pi_{title,pdatetime}((R1 - (\Pi_{postid} Comments)) \bowtie Posts)
\end{aligned}
$$