B461 Assignment 4: Relational Algebra

This assignment is designed to test your knowledge of the following lectures:

- Relational Algebra
- Joins and semijoins
- Set joins and Set semijoins in RA

For this assignment, you will need to submit 3 files. The first file is a .sql file that should contain all the SQL code relating to problems requesting the development of such code. The second file is a .txt file that should contain the output of the queries. The third file is a .pdf file that should contain your solutions for problems where RA expressions in their standard (i.e., non SQL) notation are requested. Ideally you should use latex to construct this .pdf file. Latex offers a convenient syntax to formulate RA expressions.

1 Formulating Queries in RA

1.1 Discussion

Before you solve the problems in this section, we briefly review how you can express RA expressions in SQL in a way that closely mimic their specifications as RA expressions in standard notation[1].

Consider a relation $R(A,B)$ and a relation $S(B,C)$ and consider the following RA expression F:

$$\pi_A(R) - \pi_A(\sigma_{B=1}(R \bowtie_{R.B=S.C} S))$$

Then we can write this query in SQL in a variety of ways that closely mimic its RA formulation. One way to write this RA expression in SQL is as follows:

```
SELECT DISTINCT A
FROM  R
EXCEPT
SELECT A
 FROM (SELECT DISTINCT A, B, C
            FROM        R JOIN S ON (R.B = S.C)
            WHERE  B = 1) q
```

An alternative way to write this query is to use the WITH statement of SQL.[2] To do this, we separate the RA expression F into sub-expressions as follows. (In this example, notice that each sub-expression corresponds to the application of a single RA operation. More generally, one can of course use sub-expressions that can contain multiple RA operations.)

[1] By RA expressions in standard notation, we refer to expressions that use the notations $(A : \mathbf{a})$, $\sigma...(\cdot)$, $\pi...(\cdot)$, $\cup$, $\cap$, $-$, $\times$, $\bowtie$, $\bowtie...$, $\ltimes$, and $\overline{\ltimes}$ for the RA operations. (For more detail, consult the lectures relating to RA and joins.)

[2] This is especially convenient when the RA expression is long and complicated.

Expression Name	RA expression
E_1	$\pi_A(R)$
E_2	$R \bowtie_{R.B=S.C} S$
E_3	$\sigma_{B=1}(E_2)$
E_4	$\pi_A(E_3)$
F	$E_1 - E_4$

Then we can write the following SQL query. Notice how the expressions *E*1, *E*2, *E*3, and *E*4 occur as separate queries (temporary views) in the WITH statement and that the final query gives the result for the expression *F*.[3]

```
WITH
E1 AS (SELECT DISTINCT A FROM R),
E2 AS (SELECT DISTINCT A, B, C FROM (R JOIN S ON (R.B = S.C)) e2),
E3 AS (SELECT A, B, C FROM E2 WHERE B = 1),
E4 AS (SELECT DISTINCT A FROM E3)
(SELECT A FROM E1) EXCEPT (SELECT A FROM E4);
```

In your answer to a problem, you may write the resulting RA expression with or without the WITH statement. (Your SQL query should of course closely resemble the RA expression it is aimed to express.). It should also be clear that in your solutions, you can not use the SQL set predicates [NOT] EXISTS, θ ALL, θ SOME, and [NOT] IN. You can also not use GROUP BY and aggregate functions.

[3] For better readability, I have used relational-name overloading. Sometimes, you may need to introduce new attribute names in SELECT clauses using the AS clause. Also, use DISTINCT were needed.

1.2 Problems

In the following problems, we will use the database schema that was used in Assignment 2 and Assignment3. To test your queries, you can use the `data.sql` file provided for this assignment.

Write the following queries as RA expressions in the standard RA notation.[4] In the expressions, avoid using the $\times$ operator. Rather, the use of the natural join $\bowtie$, the join $\bowtie_{\ldots}$, the semijoin $\ltimes$, and the anti semijoin $\overline{\ltimes}$ operation is encouraged. It is also not required that you "optimize" the expressions. It is sufficient that they are correct.

When you formulate your RA expressions, you can use the following abbreviations for the relations:

Relation	Permitted abbreviations
Person	P, P_1, P_2, etc
Company	C, C_1, C_2 etc
jobSkill	J, J_1, J_2 etc
Worksfor	W, W_1, W_2, etc
Knows	K, K_1, K_2, etc
PersonSkill	S, S_1, S_2, etc

Submit your RA expressions for these queries in a .pdf document. (You are strongly encouraged to use Latex.)

Then, for each such RA expression, write a SQL query (possibly using the WITH statement) that mimics this expression as discussed in Section 1.1. Submit these queries in a .sql file as usual.

Note: Do not modify anything in the data.sql file. Please use the same column names as provided. Otherwise, we will not be grading your assignments.

Question 10 is a bonus question. It is not compulsory to attempt that.

Questions:

1. Find the cname of each company that employs persons who live in Bloomington or in Indianapolis.
2. Find the pid and name of each person who (a) works for a company located in `Bloomington' and (b) knows as person who lives in `Chicago'. (Assignment 2)
3. Find each job skill that is not the job skill of any person who works for `Yahoo' or for `Netflix'.
4. Find the pid and name of each person who knows all the persons who (a) work at Netflix, (b) make at least 55000, and (c) are born after 1985.
5. Find the pairs of company names (c1; c2) such that no person who works for the company with cname c1 has a higher salary than the salaries of all persons who works for the company with cname c2.
6. Find the pid and name of each person who does not know any person who has a salary strictly above 55000.

7. Find the pid of each person who has a salary that is strictly below that of any person who has the Accounting jobskill.
8. Find the pairs (c, p) where c is the cname of a company that only employs persons who make more than 50000 and where p is the pid of a person who works at that company and who knows someone who works for IBM.
9. Find the pid and name of each person who works for IBM and who has a strictly higher salary than some other person who he or she knows and who also works for IBM.

Bonus question:

10. Find the pid of each person who has all-but-one job skill.

[4]Each of the problems relates back to a corresponding problem in Assignment 2 or in Assignment 3. You may find it useful to look at the SQL solutions for Assignment 2 and Assignment 3 as they may help you in formulating the queries as RA expressions.