Assignment 3

Continuing on our publication world, we have 5 tables containing information about scholar, publication, fields, university ranking, and conference ranking. Below we explain the semantics of the tables which will help you write the queries. The tables already exist in the sample220P.db which you will use to test your queries against. Please note that the sample file above is only a small portion of the real data. You will use the sample to try out your queries. The answers to the sample queries are provided in autograder. Please note that we will execute your queries on the full (hidden) database to test their correctness. The sample database is part of the full database.

Submission Python file:
https://drive.google.com/file/d/1_AcDmnsrGx5_0w6vO2nJL2F8xfCEBNsc/view?usp=drive_link
Autograder:
https://drive.google.com/drive/folders/1-PEL7p0cMv5yNXcfFCm0TyUquBqZhyxE?usp=sharing
Autograder Instruction:
https://docs.google.com/document/d/19SKUokd1_hqp_5WIA07rctwEbYWj7wisWRnaLtrAeXg/edit?usp=drive_link
Documentation to write relational algebra in Python:
https://docs.google.com/document/d/1UorB6g_2lo-QDg_CJUUF6mDyGFuDYT2qyeq7OhEm4H0/edit?tab=t.0#heading=h.tukp0efs6tvw
Sample Data(For you to test on Relax):
https://drive.google.com/drive/folders/1bz9Ieqeh9PwoS-r2Jrfv4v4210UOvppB?usp=sharing

Author Information Table:

This table contains the information of an author with his/her name, scholar, affiliation(university), homepage link, whether is Turing award winner, whether is an ACM fellow, region of the affiliation, country of the affiliation, begin time of affiliation, and end time of affiliation.

author('name', 'scholarid', 'affiliation', 'homepage', 'turing_award', 'acm_fellow', 'region', 'country', 'begin_time', 'end_time')

Sample Data:

Xiyuan Chen	ISK129sT	Purdue University	NULL	0	0	america	us	2022	NULL
Michael J Carey	7ahTS81	University of California, Irvine	NULL	0	1	america	us	2012	2023

You can interpret NULL in the "end_time" to mean that the author is currently still affiliated with the university. Note that in the table above, scholarid is the key. Thus, we are only storing information about the last affiliation of the scholar since each scholar has only 1 row in the table.

Also, in the table, scholar names are also unique -- that is, no two scholars have the same name. In general that would not be true, but we have cleaned the data to make things simpler.

Publication Records Table:

This table contains the summary of the publication records of each scholar in each year. The attributes are the name of the scholar, the conference abbreviation, the year of that conference, and the count of papers that the scholar published in that conference edition.

pub_info('name', 'conference', 'year', 'count')

Sample:

Sharad Mehrotra	SIGMOD	2	2024
Mike Carey	PODS	1	1992
Xiyuan Chen	ICDE	2	2006

Note that in this table, authors are referred to using their names and not their scholarid. This, however, does not cause ambiguity since remember that the scholar names are assumed to be unique. Attribute ‘conference’ is the conference abbreviation.

Field to Conference Table:

This table contains the major, which is ‘Computer Science’ in all rows for simplicity, the field under that major, and abbreviation of conference under that field.

field_conference('major', 'field', 'conference')

Sample:

Computer Science	Databases	PODS
Computer Science	Design Automation	DAC

Attribute ‘conference’ is the conference abbreviation.

Conference Ranking Table:

This table contains the ranking of each conference. It includes the abbreviation of conference, the name of conference, and the rank (A*, A, B, C)

conference_ranking('conf_abbr', 'name', 'rank')

Sample:

ICIS	International Conference on Interaction Sciences: Information Technology	A*
ACIS	Australasian Conference on Information Systems	A

University Ranking Table:

This table contains the ranking of each US university. It includes the name and alias name of the university, its state, city, zip code, its type(institutional_control), the rank(number), whether there is a tied rank, its acceptance rate, annual tuition, and number of enrollment.

usnews_university_rankings('university_name', 'alias_names', 'state', 'city', 'zip', 'institutional_control', 'rank', 'tied', 'acceptance_rate', 'tuition', 'enrollment')

Sample:

Princeton University	NJ	Princeton	8544	private	1	FALSE	6	59710	5604

Write the relational algebra query for all the questions.

Question 1

List “University of California, Berkeley” faculty who have written conference papers between 2010 - 2024 in field “Databases”(include 2010 and 2024). The answer should be in the following format.

Joe Hellerstein, SIMGOD, 2015, 2

Joe Hellerstein, VLDB, 2023, 3

John Doe, ICDE, 2017, 4

Question 2

List all conferences in “Computer Science” major where no “University of California, Irvine” faculty has ever published.

Question 3

For each “Databases” conference in each year(each conference edition), list name(s) of author and the affiliation who have published most papers.

Do not use extended relation algebra, but do this using basic operators only. If the query cannot be implemented using basic relational algebra without enhanced operators, please explain why.

Question 4

Identify names of universities, where for all “Databases” conferences, there is a publication by an author affiliated with the university between 2020 - 2024(include 2020 and 2024). For example, if an author affiliated with “University of California, Irvine” published in SIGMOD in 2021, another author published in VLDB in 2022, another author published in PODS in 2023, and yet another author published in ICDE in 2022, then University of California, Irvine should be included in the result assuming ICDE, SIGMOD, PODS and VLDB are the only “Databases” conferences in the database.

Question 5

Find all universities ranked(US News) better than “University of California, Irvine” such that the authors in those schools have never published in conferences marked A*(ICore) in the field “Databases”.

Bonus Question

The data we have provided you is real data extracted from web sites including CS ranking, US news and world report, and ICore website. In the bonus question, we would like you to describe how you would use such data to gain insights about academic organizations. In particular, how would you use such data to compare universities, in general, as well as, in a given major? Would such a comparison be different from the perspective of say a graduate student looking for admission versus, say, a recent graduate, looking for employment as a faculty? This question is open ended, and there is no right or wrong answer. We simply want to get a feel of your thought, and might select the answer we find the best to be presented at the class during lecture time (e.g., 15 minutes of the class time). Finally, when answering the question, be as precise as possible on how you will use the data/information in the tables in your comparison.

Yet another aspect (besides simply comparing universities) is to figure out “similarity” in universities. E.g., by looking at publications and the fields/conferences in which faculty from a given school publish, we could compare if schools are similar or not. Can you think of other interesting applications based on such data besides comparing schools?

Please submit the assign_expressions.py file to GradeScope and test your expressions using the given autograder. For the bonus question, you can include a detailed comment in the bottom of your python file.