Assignment 1

Due date: January 22, 2019, 11:55pm IST

General Instructions

1. Please complete this assignment *individually*, on your own.
2. You will submit 2 files: EntryNumber.sql and EntryNumber.pdf, corresponding to the dataset and a timings report, respectively.
3. Use PostgreSQL 11 for your homework. See `https://www.postgresql.org/download/` for instructions on how to download and install it on your OS. The .sql files are run automatically using the psql command using the \i option, so please ensure that there are no syntax errors in the file. *If we are unable to run your file, you get an automatic reduction to 0 marks.*

   To understand how to run many queries at once from text file, a dummy query file example.sql is available. To run example.sql in postgres, type the following command in the terminal:

   ```
   sudo -u postgres psql dbname
   ```

   and then type

   ```
   \i /address/to/example.sql
   ```

   This command will run all the queries listed in example.sql at once.
4. The format of the file should be as follows:
   - You can have a preamble section where you may create views if you like (please note that no procedures are allowed - this has to be pure sql), and correspondingly have a cleanup section where these views are removed. View names should begin with your Kerberos ID for eg. "mcs162015_view1" is a valid view name.
   - One line should identify the query number (note the two hyphens before and after the query number), followed by the actual, syntactically correct SQL query.
   - Leave a blank line after each query.

   ```
   --PREAMBLE--

   OPTIONAL DEFINITIONS

   --1--

   SQL QUERY

   --2--

   SQL QUERY

   --3--
   ```

```
SQL QUERY

--CLEANUP--

CLEANUP EVERYTHING YOU CREATED HERE
```

5. Make sure you do not use any queries like INSERT, DELETE or any meta-command in the submission file.
6. All of the queries below require an 'ORDER BY' clause. If you made an error in this clause, your answer will be evaluated as incorrect and zero marks will be awarded.
7. Errors with respect to equality and inequality conditions will also be evaluated as incorrect and zero marks will be awarded.
8. Assume set semantics, unless stated otherwise.
9. No changes are allowed in the i) data, ii) attribute names, iii) table names
10. The .pdf file should contain a bar graph. The graph should report the timings for each query of the dataset (X-axis legend is the query number, Y-axis legend is the time taken). You will need to figure out how to measure the timings.
11. The submission will be done on Moodle. Details regarding submission and auto evaluation will be updated on piazza.

1 Dataset

1.1 Instructions

1. In this assignment you will analyze publication data i.e. DBLP, the reference citation website created and maintained by Michael Ley. The analysis has to be done in postgres. We are providing you cleaned up data and you can download it from `www.cse.iitd.ac.in/~mcs172071/dataset.zip`.

   You can get more information about the DBLP from `http://dblp.uni-trier.de/xml/docu/dblpxml.pdf`.

   The data.zip file contains a tab seperated file for each table described below.(Note the order of values in file is same as attributes of table given in next bullet point). You can load the table into database from tsv file using the command -

   `copy Author from '/path/to/file/Author.tsv' DELIMITER E'\t';`
2. The database will include following five tables and you should use only these tables while writing solution of the queries. You can create temporary tables while handling any SQL query but you should include SQL queries for creating and deleting these temporary tables. Note - you don't have to define these tables in the submission file, these will already be present will evaluation.

   (a) Paper

PaperId : int	Title : text	year : int	VenueId : int

   (b) Author

AuthorId : int	name : text

   (c) PaperByAuthors

AuthorId : int	PaperId : int

   (d) Citation

Paper1Id : int	Paper2Id : int

   Note : Paper2Id is paperId of the cited paper and it is cited by Paper1Id, i.e. Paper1Id cites Paper2Id.

   (e) Venue

VenueId : int	name : text	type : text

   Note: Here 'name' is acronym of Venue and 'type' is the type of Venue like 'journals' etc.

1.2 Queries

1. For each type of venue, count the total number of publication presented at that type of venue. Your query should return a set of (publication-type, count) pairs. For example: (journals, 30000), ... Order the output columns in decreasing order of counts and break ties using publication-type alphabetically.

2. What is the average number of authors per paper? Output column: the aggregate value.

3. List the titles of all papers with more than 20 authors order by alphabetically. Output column : paper.Title.

4. List the names of all authors who never published as a single author order by alphabetically. Output column : author.name.

5. Find the top 20 authors with the largest number of publications order by first rank and on tie, order by alphabetically. Output column : author.name.

6. List the names of all authors who published as a single author more than 50 times order by alphabetically. Output column : author.name.

7. List the names of authors who have never published a paper in a journal order by alphabetically. Output column : author.name.

8. List the names of authors who have **only** journal publications order by alphabetically. Output column : author.name.

9. List the names of the authors who have published at least two papers in 2012 **and** at least three papers in 2013 order by alphabetically. Output column : author.name.

10. Find the top 20 authors with the largest number of publications in 'corr' journal order by rank and break ties using alphabetic order. Output column : author.name.

11. Name the authors that have four papers or more in 'amc' journal order by alphabetically. Output column : author.name.

12. Two major journals are 'ieicet' and 'tcs' . Find all authors who published at least 10 'ieicet' papers but never published a 'tcs' paper order by alphabetically. Output column : author.name.

13. Compute the total number of publications per year during 2004-2013, in increasing order of year. Output Column: (Year, No. of Publication).

14. How many (distinct) authors have papers with "query optimization" in the title (case insensitive) ? Output column: count.

15. Write an SQL query that lists the titles of the most cited papers order by decreasing number of citation (break ties using alphabetic order). Output column : paper.Title.

16. Write an SQL query that list the titles of all publications that have been cited more than 10 times order by alphabetically. Output column : paper.Title.

17. Write an SQL query that list the titles of all publications that are cited more than they cite by a margin of at least 10 order by alphabetically. Output column : paper.Title.

18. Write an SQL query that list the titles of papers that have never been cited order by alphabetically. Output column : paper.Title.

19. Write an SQL query that list the names of all authors who have published papers citing other papers by themselves order by alphabetically. Output column : author.name.

20. Who are the authors who published in any 'corr' journal between 2009 and 2013 but not in 'ieicet' 2009 journal? Output column: author.name order by alphabetically.

21. What is (are) the journal(s) that has (have) a strictly increasing number of papers every year from 2009 to 2013? Output column: journal.name order by alphabetically.

22. What is (are) the journal/year(s) (e.g., corr 2001) which had the largest number of papers? Output columns: journal.name, journal.year. If there are multiple answers, return all ordered by year(increasing) and break ties using journal.name alphabetically.

23. List the author that have maximum publications in each journal, sorted alphabetically by journal name. Output columns : journal.name,author.name .

24. Calculate the journal impact factor (`https://en.wikipedia.org/wiki/Impact_factor`) for each journal for year 2009. Output columns : journal.name,impact value ordered by decreasing impact value (break ties using journal name alphabetically). Note : only include those journals in the result which had at least 1 publication in the duration of previous two years(i.e. 2007-2008).

25. Calculate h-index (`https://en.wikipedia.org/wiki/H-index`) for each author. Output columns : author.name, h-index ordered by h-index value (decreasing) and break ties using author name alphabetically.