CSC4402 Project

1. The CSC4402 project is a group project. The students should form groups each of which consists of 3 to 4 people. Each group will work as an entity to produce one ﬁnal result for the project. Each group should work independently-NO collaboration between groups is allowed.
2. The implementation can be done on the class server using a relational database system MYSQL. If you would like to develop a more complex database application with GUI and Web access, you may use a different computer/database platform. However, you should inform the CSC4402 instructor about your choice of platform/project format.
3. The task for the project is to design and implement a relational database application for a suitable (real world or prototype) application domain. The project should contain the following components.
   1. Domain Application. You should understand the nature of the application, the entities important to the application, the specific attributes to be stored in the database, the relationship among the entities (and among the attributes), the constraints imposed on the database (attributes) by the nature of applications, the assumptions made by the database designer. The result of this step should be an entity-relationship model (E-R diagram) of the application, plus a list of constraints and assumptions.
   2. Database Design. You should decide the table structures of the database: How many tables should we have? What are the attributes (columns) in each table? What are the primary key and the foreign keys. Whether the tables are in some desired normal form? The theory of logical design of relational databases will be used in this phase. The result of this step should be in the form of table schemes (relational schemes) such that each table is in some desired normal form.
   3. Create the database, gather sample data and insert them to the database. Design queries and implement them using SQL to get data statistics. The SQL query should be presented, followed by the screenshot of the resulting records.
   4. Write up a report, describing the above process (a) - (c), including the script of pro-gram execution and program source code, with appropriate documentations. Finally, present the summary of the project in class. It is expected that every member of the group will participate in the presentation.
   5. The database designed should contain at least **3** entities, **2** relationships, each with a few attributes. The database should have a reasonably rich structure (e.g., one-to-many relationship) to allow interesting queries. Each table should have at least **100** records, because otherwise most of your queries will return no records. Your project should include **10** SQL "select" queries. You should design interesting queries to utilize the SQL skills (join queries, sub-queries, use of "group by" and aggregation functions, etc.)

**Choosing an application:**

Pick topics of interests, related to the course, of suitable difficulty.

Dataset

<http://webscope.sandbox.yahoo.com/> Yahoo WebScope

<http://www.data.gov/> U.S. Government's open data

<http://www.freebase.com/> Freebase

<http://www.yelp.com/developers/documentation/> Yelp

<https://developers.google.com/apis-explorer/#p/> Numerous APIs from Google (e.g., Maps, Freebase, YouTube, etc.)

<http://developer.trulia.com> Trulia <http://www.zillow.com/howto/api/APIOverview.htm> Zillow real estate listing sites

Numerous graph datasets (large and small):

<https://snap.stanford.edu> SNAP

<http://konect.uni-koblenz.de/networks/> Konect

List of lists of datasets for recommendations

<https://gist.github.com/entaroadun/1653794/>

<http://the.echonest.com/press-release/the-echo-nest-and-columbia-university-announce-million-song-dataset/> Million song dataset by Echo Nest. It contains not only the basic information of songs (artist, genre, year, length etc), but also some musical features(like tempo, pitch, key, brightness)

Movies data:

<http://developer.rottentomatoes.com> Rotten Tomatoes <http://stackoverflow.com/questions/1966503/does-imdb-provide-an-api/> IMDB

<http://grouplens.org/datasets/movielens/>

<http://archive.ics.uci.edu/ml/datasets.html/> UCI also has a collection of links to various datasets sorted for various tasks (Classification, Regression, etc)

<http://aws.amazon.com/datasets/> Amazon AWS Public Data Sets

<http://www.sigkdd.org/kddcup/index.php/> KDD Cup: annual competition in data mining, like Kaggle

**Final presentation:**

Goal and Motivation

Brief overview of the problem

Data: where you got it? Size? Format?

Your approaches

Database Design and Management (ER, normalization, etc.)

Your experimental results and Conclusions:

Querying and Data statistics

What you learn?

**Final project report:**

1. **Writeup**: Describe in depth the answers of the previous question 3a to 3e.
2. **Software**: packaging, documentation, and portability. The goal is to provide enough material, so that other people can use it. Create a zip file of the code and a short **README.txt** file. This file should describe the package in a few paragraphs, how to use it, and how to run a demo (if any).

**Submission:**

Submit to Moodle a zip file which contains the presentation (e.g. ppt), written report (e.g., word or pdf), related code (e.g., sql), table data (if the data is downloaded online, provide a link), and a README.txt.