Assignment 1: SQL Query

1st Version Due: Jan 15, 2015
Revised Version Due: Jan 22, 2015

Project: Create a MySQL query that asks an interesting question about a database

The purpose of the assignment is to creatively look into a database to result in interesting and/or unexpected discoveries (KDD: Knowledge discovery in databases)

Database Resource:
The database to be used is the list of checkouts from the Seattle Public Library retrieved by the "Making Visible the Invisible" project. There are appx 30,000 checkouts per day, 10 million per year, currently over 77 million data entries.

Multivariate Datasets include:
. ItemNumber: Collection acquisition time-stamp
. bibNumber: Each topic-specific item in collection
. Barcode: Each item has a unique rfid sticker
. Check-out/check-in hour/day: In/out interaction with database
. ItemType: books, cds, dvds, music sheets, etc.
. Title
. Dewey Classification (all non-fictional subjects)
. Subjects: Keywords

Access the Database with MySQL Workbench:
Host: tango.mat.ucsb.edu
User: mat259
Password: V1sual1zat1on
Database: (optional)
Port: 3306

Approach:
. Get an overview of the dataset
. Explore topic specifics of personal interest
. Embrace detail to integrate complexity

What to look for:
You can say something about the cultural content of the database, about your own interests through how you explore the database, or analyze the peculiarities of the structure of the database

. Search for (unexpected) patterns in the data
. How often something happens (frequency)
. Anomalies in the system (errors, outliers, etc.) http://en.wikipedia.org/wiki/Anomaly_detection
. Association: Search for relationships between variables
. Do statistical analysis using MySQL aggregate functions:
http://dev.mysql.com/doc/refman/5.6/en/group-by-functions.html

What are the conditions:
. Results to consist of multivariate data to allow for minimum 3 columns metadata
. Standardize columns for output to .csv file: (vertical, horizontal, pixel value + more)
. Your search results should be sufficient to feature subtleties in the data to be expressed visually within a screen size between 1920x 1280 to 2560 x 1850 pixels

Process:
. Imagine a question
. Explore all fields of the database
. Test the query for logical and other errors
. Start with Booleans (AND, OR NOT), wildcards
. Make the query efficient (finetune)
. Interpret the results (data analysis)
. Export to .csv file to be used in visualization

Create a Project Report (to include):
. Concept/Question (describe what question you are exploring)
. Provide the Query
. Explain the Query
. Provide the results
. Give Processing Time
. Give an Analysis

Post your Project:
. Print out as pdf.
. Post your report at http://www.mat.ucsb.edu/forum/ at MAT 259/Winter 2015
. PostReply to Proj 1: Data Query with a short abstract of your pdf
. Attach the pdf.

Grading
. Standard Completion of Project: B
. Revised, advanced functions: B+, A-
. Innovative question: A

Previous Student Examples from last year's student forum:
Standard Examples:
http://www.mat.ucsb.edu/forum/viewtopic.php?f=65&t=241&start=10#p1507 (Rob Miller)
http://www.mat.ucsb.edu/forum/viewtopic.php?f=65&t=241&start=10#p1508 (Laks)

Advanced Examples:
http://www.mat.ucsb.edu/forum/viewtopic.php?f=65&t=241&sid=ef537f62a834c0984ef326f6fd030f79#p1499 (Grant McKenzie)
http://www.mat.ucsb.edu/forum/viewtopic.php?f=65&t=241&sid=ef537f62a834c0984ef326f6fd030f79#p1505 (Kitty Currier)

Term Definitions:
. Clustering: http://en.wikipedia.org/wiki/Cluster_analysis
. Classification: Organize the data http://en.wikipedia.org/wiki/Statistical_classification
. Frequency Pattern: http://en.wikipedia.org/wiki/Association_rule_learning
. KDD: http://en.wikipedia.org/wiki/Data_mining