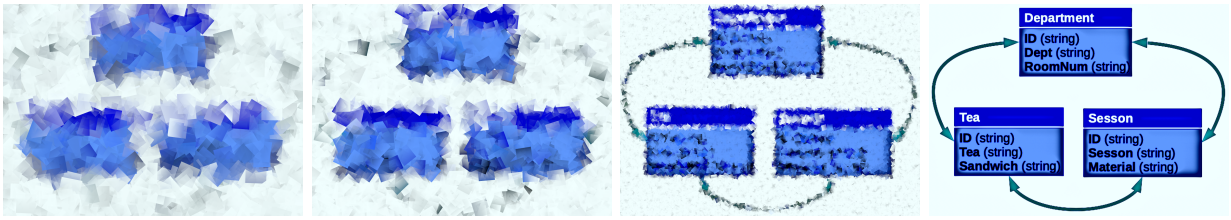
CMPSC 312
Database Systems
Fall 2020

Final Project Assignment:
Advanced Topics in Data Management
Place the buildfiles, data and related technical content in `src/`
and the *statusUpdate.md* and *report.md* in the `writing/` directories.



Objectives

To produce a project where databases are integral part of the design and implementation. To gain experience working with real-world applications where database systems are used.

GitHub Starter Link

https://classroom.github.com/a/SMTIG73x

To use this link, please follow the steps below.

- Click on the link and accept the assignment
- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab,
- Clone this repository (bearing your name) and work locally
- As you are working on your lab, you are to commit and push regularly. The commands are the following.
  - `git add -A`
  - `git commit -m ‘‘Your notes about commit here’’`
  - `git push`

Introduction

Throughout the semester, we learned much about the construction and usages of databases. During our class and lab time, we discussed concepts, theoretical mechanisms, management and general technologies which bring databases to life in academia and industry. The topics were applied to structured query language (SQL), the implementation of database applications (i.e., using Python

and Django), NoSQL systems (MongoDB) and a visual database (Neo4j). All of these systems were designed to enable users to locate meaningful information from collections of data.

During this work, you learned the differences between database types for different applications. In addition, you gained a deeper understanding of what types of data (or formats of data) may be used by each database system so that your applications were able to be efficiently designed, maintained and run. Such topics in our course included; integrity constraints, attributes, table dimensions, queries, documents, collections, nodes and edges and others. In this final project, you are to combine all your knowledge of database construction, with a research question or specific application, to create an efficient database that will be able organize the data of your research.

Creating a database to answer research questions

Your project will concern responding to research questions which you will use a database to answer. **Choose a project which presents several questions for which the data you collect will allow you to answer.** Remember, that this assignment is not so much about being able to analyse data to resolve some issue, but rather, it is to showcase your ability to organize collected data so as to be in the position to perform such an analysis. Your project is therefore to respond to your research questions by creating a database system to address the necessary queries to understand or resolve the questions that you ask of your data.

Choose your data: This project invites you to sufficiently explore an advanced topic in databases which will be helpful to storing and ordering your data for a research project. You will collect data to be organized in a database that you create and then use queries to answer research questions. This data must come from public databases where it is not artificial. For example, data may be found at the Allegheny College Library Scholar Databases (found at the following address: `https://allegheny.libguides.com/az.php`) where you can connect to online sites such as the *World Health Organization*, `http://www.who.int/gho/en/`, *UniProt* `http://www.uniprot.org/`, *ProQuest's Biological Science Collection* `https://search.proquest.com/biologicalscience1` and similar types of resources. Off-campus resources include *Kaggle's Datasets*, `https://www.kaggle.com/datasets`, *Data.gov*, `https://www.data.gov/` and others which can be found using online searches.

Choose five research questions from the data: From your data, you are to find five (5) *intelligent and meaningful* research questions that can be answered using queries. Remember, your work in queries should be sufficient to determine responses to your five research questions.

Choose your database technology: We have worked with several different technologies this semester and each type has specific advantages that you could use. For instance, if you are studying potential relationships, you might choose to work with Neo4j, and if your data is messy, then perhaps MongoDB would be more fitting. If your data can be organized into tables, then perhaps SQlite would be a better selection. There is much overlap between each database and your choice of database technology might actually be based on your own desire to work with one system over another.

Write your report: Your project should result in a detailed report written in Markdown and contain graphics, plots, snippets of code or whatever you require to make your research project abundantly clear to the reader. The report should include a description of the database that you

choose for the project, the data, the five research questions, the queries used to respond to the questions and the conclusions that you draw from the queries. Please be sure to add citations of your data in this report so that your work can be followed and reproduced. In addition, you will want to include any details of how you built your database, assembled the datasets and what conclusions you have drawn from your queries.

Summery of Deliverables

You are to submit your report, code and data for your final project deliverables. The written report should be precise, formal, appropriately formatted, grammatically correct, informative, and interesting. The source code that you write to build your database must be carefully documented and tested. If you install and use a data management framework, the steps for installation and use should be clearly documented. Your data should be the cleaned version. Please note: if your data set is over 5 megabytes, please consider compressing it into a zip file for your submission.

Your project (and its deliverables) must conform to the following guidelines.

- **Report**: You are to submit a report document that details the project, including the citations of the data and any other elements used to complete the work. The five (related) research questions are to be discussed along with the methods (i.e., data, software packages and etc.) that you utilized to address them. Your report will contain the queries, screen shots, plots and other visual tools with discussion to explain your conclusions. Some of these details are explored below. In addition, you will want to address the database system, including the software version, how you loaded the data and the database's schema (as necessary) to explain why you chose this software.

  Please note that this is a technical document and will have the same types of sections as other technical documents that you have come to know during your exploration in the literature. In the past, final project reports from this class contained about 10 to 15 pages (text, technicals and graphics). Below are explanations of some of the areas of discussion of the report.

  1. **Data**: You are to use real data which you have either, amassed from experimental techniques (i.e., you acquire your data by your own designs), or has been obtained from public repositories (mentioned above) where the data has been collected from real events.
  2. **Database software**: Your project must have a justification for the software packages that you chose to use. You may use any software that we have discussed in class (i.e., SQlite, Python, Django, SQLBrowser, MongoDB, Neo4j) or any relevant software package that you are interested in pursuing on your own.
  3. **Source Code**: Any code must be submitted for the instructor to evaluate. Please include comments and print statements, where necessary, to help with reading the code and to provide details of its execution.
  4. **Queries:** In your project, you will be using this database to extract meaningful information (patterns) from your data to explain some phenomenon. Exactly which patterns to find will be left to you and your group. However, these patterns must be well thought-out and relevant to the theme of your project. For this task, queries must be used and each query is to be outlined with a short justification about why the particular query was used in your project.

5. **Conclusions from the Queries**: Your written project must reiterate the research question and give your formalized conclusion which is based off your queries of the database containing your data. Please be sure to justify your conclusions.
6. **References**: Please be sure to cite your data, facts or other important addition to your work. In addition, each claim that you make which is not necessarily public knowledge, you ought to have some reference to a primary, peer-reviewed article, to support your statement. If you are mentioning a novel tool, please search for its supporting article to add to your reference section and tie this reference to your discussion of the tool.

Summary of the Required Deliverables

This assignment invites you to submit an electronic version of the following deliverables through your group GitHub assignment repository. Do not forget to add your name to your work.

1. **Status Update Due**: Wednesday, 2nd December 2020, 11:59pm. The file `writing/statusUpdate.md` will contain your writing about the following details. This work will cover about half a page.

   (a) The title and subject of your project
   (b) The main questions are you that you are addressing,
   (c) The citations of your data and some leading references to support the project
   (d) The steps that you have already taken
   (e) The step left to complete the project

   Please be specific and provide a complete list names of all group members in this document so that your instructor can give credit for this work.

2. **Final Project**: The file `writing/report.md` will contain the written portion of your deliverable. Your report is to contain the above-mentioned details and items. See the above discussion for details of the content. You will submit all work using GitHub where you have placed all relevant files and with the report.

   Note the deadline in the header of this document.

3. **Source**: In the directory `src/`, you are to add all code and data for your project is to be submitted. In your report, please include instructions about how to run your code with the data.

In adherence to the Honor Code, students should complete this assignment individually. While it is appropriate for students in this class to have high-level conversations about the assignment with other class members, it is necessary to distinguish carefully between an individual who discusses the principles underlying a problem with others and the student who produces an assignment that is identical to, or merely a variation on, the work of someone else. As such, deliverables that are nearly identical to the work of others will be taken as evidence of violating the Honor Code. Students should contact the course instructor with questions about this course policy.