ECE 356 Project

COVID-19 Data

This document is a description of details relevant to those doing the course project using the **COVID-19** datasets. It provides details about the dataset, as well as suggestions pertaining to the client application, the entity-relationship design, and the data-mining exercise.

Data Source

There is so much data surrounding COVID that it is not easy to pick any clear suitable source(s). Therefore groups that are doing COVID projects can do so by using selected data from:

`https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge`

and/or selected data from StatsCan mortality data in general and COVID-specific mortality data. For example:

`https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=1310039201`
`https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310077501`

but there is also a wealth of other relevant COVID data from StatsCan.

We have made the CSV files from Kaggle available on marmoset04 in "`/var/lib/mysql-files/06-COVID/`" using exactly the names as on the Kaggle site. We have not put CSV data from StatsCan in the 06-COVID directory as there are different options for downloading it, and it is likely that you will want to think about how best to access the data. In addition, the StatsCan data tends not to be in simple CSV files but has a more complex format, making direct loading non-trivial. If you are using this data you may, therefore, use the scripting langauge of your choice to pre-process the data and then load it into the database server. If at any point you want to do a direct load of CSV files into the server, please let the relevant instruction-team member know and we will make arrangements to upload the file to /var/lib/mysql-files.

For the CORD data on Kaggle, you should look on Kaggle to determine what the different attributes are within the dataset.

Client Application

As noted in the main project document, there will be little additional to add to the generic client-application requirements listed there. If you want a sense of what a client application for the movie domain should do, you should think about the potential users of such a database. In the case of COVID data there is a vast vange of possible users, and so only very generic advice can be given here. Users range from governments wanting insights into the effects of policies on hospitalizations and deaths, the economy, the effects on other aspects of people's lives (*e.g.*, the mental-health effects of various policies), *etc.*, charities wanting insights into the effects of the situation on community needs, *etc.* Broadly, any user will have two things in common: the need to add data to the database as it becomes available and the need to look at the data in a way that is relevant to the user. You should define this aspect of your project first, since it will inform all other aspects of your design.

Entity-Relationship Design

Per the main project document, you will need to determine an appropriate ER design for your dataset. There have been no prior projects done using this data, nor is it reasonably possible to define what appropriate entity sets might be in the absence of a clear notion of what specific user/client-application your group has in mind. As such, if you have difficulty in thinking about different relevant entity sets for this domain you should consult with your designated instruction-team member. If you have only three or four entity sets, with just a few relationship sets between them, and only couple of dozen attributes, your project is probably not a top quality project. The NHL database should give you a sense of the size and scope of what is expected.

Data-Mining Investigation

For COVID data there are innumerable possible data mining-exercises that are worth considering. Unquestionably, one of the most useful ones is expressed in the question, "What factors (attributes) best predict mortality?" This is likely best approached as a classification exercise, where you have the outcomes as death/non-death, and then build a classification tree on the data. Other questions of relevance include, "What policies lead to the best tradeoff between the economic consequences and the mortality consequences?" and "Is it possible to generate a clustering that suggests connections between mortality and other factors?"

If you have difficulty thinking about an appropriate data-mining exercise, you should consult with your designated instruction-team member. If you think you have a good idea for a data-mining exercise, it is probably worthwhile checking with your designated instruction-team member to confirm that it is of appropriate scope.