ECE 356 Project

2020 Election Data

This document is a description of details relevant to those doing the course project using the **2020 election** 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 are several sources of data for those doing a project in the 2020 election domain:

`https://www.kaggle.com/manchunhui/us-election-2020-tweets`
`https://www.kaggle.com/etsc9287/2020-general-election-polls`
`https://www.kaggle.com/unanimad/us-election-2020`

The first comprises two CSVs containing tweets from 20thOctober to 8thNovember with 21 attributes in each. The second source contains three files with polling data (for both 2020 and 2016 so as to provide context) and detailed per-county data. The third source contains detailed results of the election. You need only concern yourself with the president_... files, with the president_county being the most relevant. All CSV files from the above sources have been made available on marmoset04 in "`/var/lib/mysql-files/05-2020-Election/`" using exactly the names as on the Kaggle site.

The tweets include user location information which you should be able to use to place the user within a county. You may need to use the USAddresses data for that, which is available at:

`https://www.kaggle.com/search?q=openaddresses+in%3Adatasets`

and the relevant CSVs are on marmoset04 in "`/var/lib/mysql-files/USAddresses/`" with separate CSV files for each state, by two-character state code. For example, Texas addresses are contained within the `tx.csv` file, California ones in `ca.csv`, *etc.*

You may also find the data within

`https://www.kaggle.com/johnjdavisiv/us-counties-covid19-weather-sociohealth-data`

useful as it contains per-county information which may be relevant for your data-mining exercise, though this is entirely optional.

In all instances, you should look at any of the relevant datasets on Kaggle to determine what the different attributes are within the CSV.

Client Application

As noted in the main project document, there will be little additional to add to the generic client-application requirements listed there. The most likely user for an election database would be political parties attempting to understand how people voted and where votes changed. In that regard, the client application should allow a user to find out information about various counties, including getting a sense of how many tweets there were from a particular county and how that might connect to voting results, whether the tweets with a Trump or Biden hashtag were positive or negative about the candidate, *etc.* It is likely that a user might want to also make annotations about a particular county, and so some form of ability to add such data to the database should be possible within the client. Data mining the results would likely be central within such a client, also, and you might want to allow a user to perform data mining operations over a particular segment of the data. For example, you might want to allow the user to classify counties by data. See below.

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 in this course using this data. That said, there are certain entity sets that should be apparent within the data, such as "county", "state", "tweet" and likely several others. If you have difficulty in thinking about different relevant entity sets for this domain you should consult with your designated instruction-team member.

Data-Mining Investigation

For election data there are numerous possible data mining-exercises that are worth considering. One of the most useful ones is expressed in the question, "What factors predict if a county will vote Trump *vs.* Biden?" A classification approach would be very useful here, as the result data is contained within the database, and so you can use the standard classification-tree building algorithms over the data. Other questions of relevance include, "What are the factors that determined changes in polling data between 2016 and 2020?"

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.