Team Corybantics
Kyle Kimery and Ivy Markwell
12/6/16

CS327E Elements of Databases Final Report

Team Corybantics: Kyle Kimery and Ivy Markwell

1. Introduction

Team Corybantics undertook all of the assignments and projects in CS327E with two objectives: learn more about working with databases, and learn more about a topic that interests us. In both regards we have succeeded. At the start of the semester, with these objectives in mind, we choose endangered wildlife as our topic of study. This decision was perhaps the most important and difficult of the entire class.

2.0 Motivation

To meet our personal objectives we needed a dataset that was simple enough to understand and work with, but not so simple that the results of our project were meaningless and trivial. Additionally, we wanted a dataset that interested both of us and would keep us motivated and intellectually stimulated for the several months that we would be involved with the data. The decision to research endangered wildlife was a compromise after many hours of discussion and fruitless investigation of other options.

Other topics we considered included: election results by county, video game review scores, and data from epidemics. In each case we faced problems that disqualified that option from consideration. We were unable to find reliable, trustworthy data for video game scores. We found an overwhelming

amount of data for epidemics, to the point that we weren't able to understand what exactly we were being shown in the pages and pages of data we found. For political results we found excellent, understandable data, but ultimately weren't particularly interested in dedicating our efforts for the next few months towards the topic. Endangered wildlife interested us, and the data came from a very reliable source, but we had second thoughts about how simple the data was. We worried that any analysis on our dataset wouldn't produce particularly interesting insight.

While working on the final project, we incorporated data from Twitter along with our original endangered animal data. The goal of the data from Twitter is to highlight any relationships between the amount of "attention" (quantified in terms of retweets, likes, shares, etc) the public gives to wildlife, nature and endangered animals in a given area.

Thus, the ultimate "goal" of our project is to explore the links between the number of endangered animals in a country and the public's level of interest or concern in the endangered animals that exist in that country. The hypothesis we are investigating is that increased publicity for endangered animals helps save them from extinction. This can be tested over the long term by studying the Twitter-interest in a country's animals with the number of animals that are on the threatened and endangered list. If in 5-10 years (if Twitter exists) we would expect to see a decline in the number of endangered animals in these high-concern areas or at least the downgrading certain animals' endangered status.

Team Corybantics
Kyle Kimery and Ivy Markwell
12/6/16

2.1 Data Modeling

As we began work on modeling our data, some our earlier fears about our dataset were realized. The data was so simple that we struggled to meet the required number of entities in our data model. We created tables modeling both mammals and birds, but functionally there was no difference between the two and making a distinction between the two was pointless. The data didn't have any attributes for birds that it didn't have for mammals and vise versa. To remedy this, Ivy had the idea of locating additional data that would create a distinction between mammals and birds so that we could justify storing the data in separate tables. Thus, we added an attribute to bird to show whether or not they were migratory.

The next hurdle we faced with our data modeling was the inconsistency in our data for the location of each endangered animal. Some of the animals only had their continent listed, while others were as specific as individual counties in states in the US. We needed to find a way to store the areas for each of the animals, without losing too much information or causing gaps in our data. The solution we came up with was to "zoom out" and use the home continent for each animal. We had a disappointing loss of specificity, but otherwise we would have giant holes in our data for the home range for those animals which had very general home ranges.

The final hurdle we faced for data modeling in Lab 1 was that many of the animals existed in multiple continents. The solution was to create a junction table "Animal Range" between our

Endangered Animal table and our Continent table. Conceptually we struggled to understand what this table would represent in the "real" world, but it was the only feasible way to make our data workable.

While modeling data for the final project, we encountered few difficulties. We created an entity to represent a tweet and used the tweet id as the primary key. To track public interest in a tweet, we added the number of retweets a tweet has as an attribute. Thus we are using retweets as an abstraction of public concern. There are obvious problems with this and it is not a direct relationship, but it is a simple and elegant solution to our problem.

2.2 Schema Design

While creating our tables and attributes in MySQL we experienced challenges with the order we created tables and with what data types should represent each attribute. Our Data Model calls for near-incestuous parent-child relationships and navigating these relationships was perplexing. We had to create the tables in the correct order or else we would be inundated with errors. Finding the correct order required a thorough understanding of not only MySQL but also our own data and how everything related to everything else.

Determining which data types to use to represent each attribute was another challenge we encountered. Neither of us had encountered varchars in our combined programming experience, so the idea was a bit strange. After doing research on our own and looking at example code we found online, we elected to set the length of the varchars at 255. We didn't want to place arbitrary constraints on

possible data. Although we could have simply found the longest names and text strings in our dataset and set the varchar length accordingly, we would be limiting all future data as well. Although this project isn't necessarily looking so far in the future, we decided any potential efficiency loss (although our research suggested there wouldn't be any) was worth building the habit of future-proofing our code.

We were also a bit intimidated by the sheer number of data type options we were afforded. MySQL offers a wide breadth of data types to choose from, but our choices were safe, simple and boring. It may have been more rewarding in the long-run to have experimented with other data types and strayed from our comfort zone a bit more.

On the final project, our schema did not change much. We added a new entity for the Tweet table, which connected nicely to our existing table for Continents. Our Tweet attributes were simply components of the tweets, stored as ints and varchars. The biggest change was the inclusion of the JSON data in the Tweet entity. Understanding JSON and how MySQL works alongside it was key to accessing the cornucopia of data that is Twitter. Using this datatype was a departure from the norm for team Corybantics, and we were forced to go back over our notes and readings in order to utilize JSON effectively.

Team Corybantics
Kyle Kimery and Ivy Markwell
12/6/16

2.3 Data Collection and Loading

Collecting and loading our data proved to be a considerable challenge at certain points. Our dataset was inconsistent and this created a number of headaches when trying to populate our tables with python and pymysql. As previously mentioned, the home ranges of the animals varied widely in scope and scale. The only solution we could come up with was to manually change each offending data point. This involved a team member looking up each county, country, region, etc to find out which continent it belongs to. We attempted to automate this process in our import scripts, but we found that the data varied so wildly that it was faster to do everything by hand. This was an excruciatingly slow and inefficient process, and in the future we will make certain our data is internally consistent so that we can avoid this painstaking manual data input.

Another issue we discovered was that there were duplicate entries in the csv file. The primary key of our Endangered Animal table is the scientific name of the animal, a unique and non-changing value. However, in a few instances, our data set had duplicate entries for scientific names. We double checked, and the problem was in fact with the data itself and not with our scripts. This obstacle was overcome by manually deleting the (few) offending data points to preserve uniqueness of our primary keys.

Both of the above issues stem from our poor choice of a dataset. Although the source of the data is trusted, the format that the data was stored in was shoddy. Duplicate entries and wild inconsistency in the dataset cost us hours of manual work and decreased the usefulness of our database. In the future,

we will make certain that every aspect of the data in a dataset is to our liking and workable. Simple problems can be solved in our scripts, but fundamentally broken data cripples our effectiveness and efficiency.

When collecting Twitter data for our final project, we encountered other errors. Our initial idea for using the tweets was to connect the Tweet entity to Endangered Animal, rather than to Continent. We would be observing the same relationship between public interest and endangered status, but on an animal-by-animal basis rather than continent-by-continent. Unfortunately we were forced to abandon this idea. Our initial tests of our twitter api script implementing this idea triggered Twitter's timeout feature. We were trying to query the site once for each animal in our database. All five hundred of them. We were limited in our querying because we are on the most basic (and free) access to Twitter. On top of being forced to wait, we realized that our queries were much too specific and were returning no results.

To reduce the number of queries while increasing the amount of data returned, we decided to broaden the scope of our Twitter queries. Our solution was to search for mentions of terms such as "#endangered" and mentions of data from our Continent table. The result is data showing tweets about endangered animals in each country. In this instance, we were forced to compromise on our vision of our database due to our limited resources. The result, while different from what we originally intended, provides a different look at the same data and is still enlightening.

Team Corybantics
Kyle Kimery and Ivy Markwell
12/6/16

2.4 Queries

Constructing the queries is where the simplicity of our dataset came back to haunt us again. From the offset we knew that our data would limit us, but we had hoped that we would still be able to locate something interesting to analyze in our results. Our schema involves a small number of entities which in turn have a small number of attributes. The requirements of the lab asked us to provide many queries, but our data didn't really allow for any meaningful searches. We struggled to come up with any query that wasn't a simple variation of counting the number of rows in a join. It was interesting to compare the number of endangered animals per continent, but there's not enough substance there to justify making a dozen queries into it.

Using user-input in our query interface was a way for us to make our searches a bit more engaging to the end-user. Despite this, we were still plagued with questions of "What is the point of this query?" and "Why should I care?". At the conclusion of Lab 3, it felt as though we hadn't accomplished anything useful with our work other than learning SQL.

Upon reflection, the only ways we could have fixed the fundamental uselessness of our database was to either scratch the entire thing and start from the ground up with a dataset with more interesting and diverse attributes and tables or we could have added in other data from a new dataset that tied into our endangered wildlife database. Thankfully, the final project asked for just that.

By incorporating data from Twitter, we were able to overcome our database's existential crisis. With retweet data about endangered animals tied to continents, we are able to craft queries for our user that demonstrate more than the simplest, surface-level relationships between our entities. It also reignited the motivation for this semester-long project.

2.5 Challenges

As has been alluded to many times in this report, the greatest challenge we faced over the course of this semester has been selecting our dataset. We knew that this decision would shape the rest of the project for us, and yet we still failed properly vet our data before committing to it. The primary shortcomings of our data were: simplicity, inconsistency, and narrowness. The data was too basic, averaging a single attribute per entity, to allow for "deep" searches. The data was too narrow, being forced to limit our animal habitats to just continents effectively erased the massive breadth of data that existed for some, but not all, the animals. And the data was inconsistent, forcing us to waste time manually changing values.

3. Conclusion

This semester would have been significantly easier for us if we had elected to research another topic. There were plenty of datasets perfectly prepared for our purposes. Perhaps we could have pursued one of them, but what would be the point? Endangered wildlife is something that intrigued

both of us, and working on this data is more enjoyable than working on an easier dataset, because the topic is something that my partner and I truly care about.

In the long run, this experience may be more valuable because of the obstacles we faced this semester. While it would have been nice to work with the perfect dataset, that is a situation that is likely rare in the “real world”. There will be times in all fields and disciplines where we are forced to troubleshoot, change plans, and creatively approach difficult problems. Thus we leave this class not simply with a basic understanding of SQL and relational databases, but with hands-on experience working with challenging professional projects in less-than-ideal situations. I am reminded at this time of Robert Frost’s famous words: "Two roads diverged in a wood, and I- I took the one less traveled by, And that has made all the difference.”