Group 6 - Phase 0

Kevin Dombrosky
Rochester Institute of Technology Student
610 Park Point Dr
Rochester, New York 14623
kfd6490@rit.edu

Brittany Purcell
Rochester Institute of Technology Student
107 Weldon Street
Rochester, New York 14611
blp6903@rit.edu

ABSTRACT

This paper discusses the outline of the group project for Group 6 in the Intro to Big Data class at Rochester Institute of Technology, CSCI-620. The project requires use of a database management component and a database analytics component. The paper further breaks down the dataset, database management system, and application that will be used for the project's completion.

Keywords

Phase 0, Group 6, Plants

1. INTRODUCTION

The group project for CSCI-620, Intro to Big Data at Rochester Institute of Technology, requires students to explore storage and analytics with a large data set.

There are two requirements for this project. First, it must include a database management system (DBMS) that is preferred to be relational in nature. This includes data modeling and Entity-Relationship Diagrams, SQL programming, and relational design practices including indexing and constraints within the DBMS. Second, some form of database analytics. This can include a visualization of the data, or showing use of a data mining technique, including classification, clustering, and association.

These requirements are planned to be accomplished by Group 6 by creating an application that shows the geographical location of various types of plants in the United States and Canada using the Plants dataset from the UCI Machine Learning Repository. The database system is intended to be the relational database, MySQL. The analytics component is intended to be a visualization of what states and provinces have each plant within them.

2. BODY

2.1 Dataset

The dataset that is to be used by the application is the Plants dataset from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Plants). It includes over 22,000 records of scientific names of plants and a comma separated list of state abbreviations of what states the plant can be found in. To make the data easier to ingest and comprehend, a file was included for converting the abbreviations for states and provinces.

The already present structure in the dataset should cause for relatively easy data ingestion. The group may need to express caution when character encodings are considered, since several special characters appear present in the data set. Ingestion will likely be carried out by a script outputting the SQL statement to insert the records into the database.

2.2 Database Management System

The database management system that the team intends to use to store the Plants dataset is the MySQL DBMS. MySQL presents itself as the most viable option for the team. The barrier of cost is the largest consideration for the team, and MySQL being free makes it very appealing among other systems. From the subset of databases remaining in consideration, the team has the most experience with MySQL, so the team wants to complete the project using MySQL as their relational database management system.

Another database system that was considered for the project was PostgreSQL. While PostgreSQL offers great customization, it is heavier duty than what is required and the team collectively has less experience working with PostgreSQL.

2.3 Application

The application design being considered by the team is simple, but also still in a relatively conceptual stage.

The application will primarily feature a search box and a map of Canada and the United States. The search box will be expected to allow for searching of whole strings of a plant's scientific name. If a whole string is input, the map should be highlighted as to what locations the plant can be found in.

There are additional features planned for the application, time permitting. One of these features are for the search box to allow for partial string searching of plants scientific names. Doing so would bring up a listing of plant names that the user can the select from to highlight on the map. Another feature that is desired is to incorporate common names of plants to be able to be searched. However, the common names of plants is not included with the data set, so this will likely require manual data collection and input to the DBMS, or reaching out to another dataset. Another feature that was brainstormed by the team is to allow for the user to click on a state or province region to bring up a list of plants that are located in that region.

These additional features are desired in the end product and will be pursued. The barrier that the team is concerned about limiting their time for the project is simply inexperience with certain technologies. The team hasn't worked with any packages to allow for visualization before. This may be able to be circumvented by instead creating a web application.

3. CONCLUSIONS

The team plans to design and implement an application that uses the Plants dataset from the UCI Machine Learning Repository. This application will use a MySQL relational database management system to store the data. The team has high hopes for the features of the application, time permitting, and are looking forward to the challenge of learning new technologies regarding visualization.