CSCE100 Introduction to Informatics
Fall 2018

Final Project Assignment (Group): Informatics on Chronicling America's Repository of Historical Newspapers

Points: 250 points. Assignment Date: October 11, 2018 Due Date: December 03, 2018

Objectives

1. To define and describe a problem in informatics focused on the Chronicling America's repository of historical newspapers
2. To carry out the various steps of Informatics to solve the problem: data preparation, data cleaning, data pre-processing, data analysis, and data visualization
3. To utilize MySQL queries to *explore* data provided in the repository
4. *Data Preparation*: To utilize a sequence of MySQL queries to extract data necessary for your Informatics problem
5. *Data Cleaning and Data Pre-Processing*: To implement programs in Python to clean the data and pre-process data as needed
6. *Data Analysis*: To implement programs in Python to analyze the data to generate information or findings (e.g., trends, patterns, or other useful insights)
7. *Data Visualization*: To visualize the information and findings clearly and intuitively for others to comprehend the information and findings by producing infographics
8. To learn about teamwork and work as a team

Additional Resources

Official MySQL Tutorial: https://www.tutorialspoint.com/mysql/index.htm
SQL Tutorial: https://www.w3schools.com/sql/

Database

Referring back to Programming Assignment 4, a database has been pre-built for you. For this Assignment, you are required to access the database using MySQL queries to understand better the database. The database is a subset of the Chronicling America's repository, which is maintained by the Library of Congress. Each entry in the database is the metadata (attributes) of a newspaper page that has been archived and scanned into a digital form. The database is made up of multiple tables, with each table capturing parts of the data. The relationship diagram of the tables is shown in Figure 1 below.

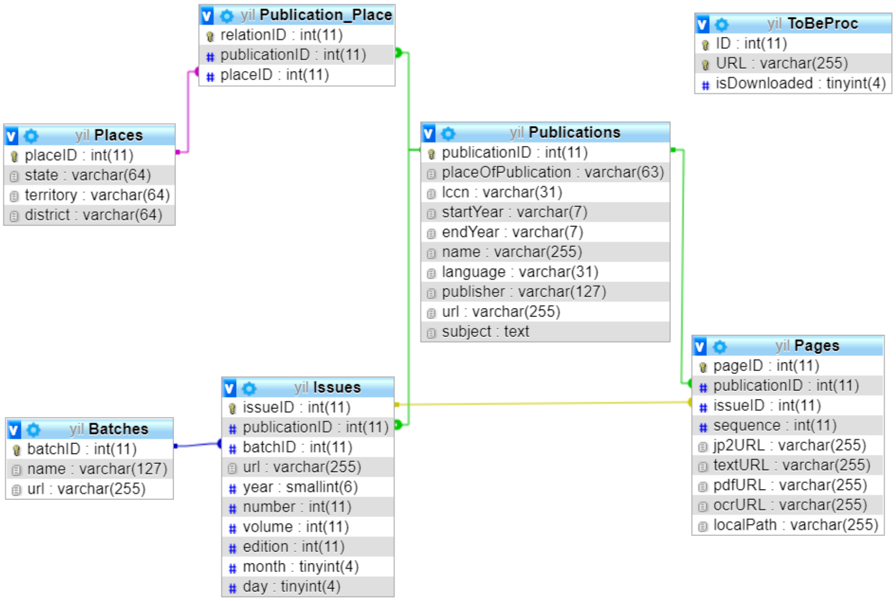

Figure 1. Entity-relationship (ER) diagram of the 7 tables used in the database. Take for example, the "Issues" table, it has a common link to the attribute "publicationID" with the "Publication_Place" table, and a common link to the attribute "batchID" with the "Batches" table.

Final Project Problem

The U.S. Library of Congress's Chronicling America repository (https://chroniclingamerica.loc.gov) is a collection of America's historic newspaper pages from 1789-1963. We have downloaded 8000+ images and each image's metadata. The database schema shown in Figure 1 contains the metadata. For this final project, identify a specific problem in Informatics using the database. For example, are there interesting questions to answer? Are there patterns that you want to find? Are there trends to be observed? Explore the database to get yourself familiarized with the data first, and then form your questions that will drive your Informatics efforts.

In addition to the above database, we also have the actual document images (8000+) at our local directory. Interested teams can request access to these images and we will also provide image processing tools to extract information from these images. Thus, another type of informatics problems that a team can perform is based on the actual images, instead of the metadata. Of course, yet another type of informatics problems is combining both data derived from images and the metadata (such as correlation analysis to associate noisy images with certain publications or publishing companies, or publication years).

Requirements and Handin

1. The submission deadline for all handins is December 03, 2018, 9:30 AM, which is also our Demo Day. **Late handins will *not* be accepted or graded.**
2. **Final Project Proposal Q&A** (0 points but REQUIRED):
   a. **November 19, 2018:** Each team is required to turn in a 3-page Final Project Proposal and present their proposal (using Powerpoint slides) in class and participate in a Q&A discussion.

b. **The Proposal must include at least the following:** (1) a Team Name, (2) The names of all team members, (3) Proposed informatics problem, (4) Proposed tasks to solve the informatics problem, and (5) Specific tasks and responsibilities for each team member.

3. **Final Project Report** (100 points): Each team is required to handin a final project report. The following sections are required:

   a. **Introduction** (15 points): This section describes your informatics problem

   b. **Data Preparation** (15 points): This section describes your data exploration strategy and how you extract with the data that you used for your final project. You must include the exact MySQL queries used in this section.

   c. **Data Cleaning & Pre-Processing** (10 points): This section describes your data cleaning and pre-processing strategies used. Must refer to any Python programs that you used to accomplish this step.

   d. **Data Analysis** (20 points): This section describes your data analysis solution. Must justify your analysis (similar to how you justified the statistics used in your Programming Assignment #3). Must refer to any Python programs that you used to accomplish this step.

   e. **Data Visualization** (20 points): This section describes your data visualization strategies. Must justify your strategies. Must refer to any Python program that you used to accomplish this step. Must also provide the infographic.

   f. **Conclusions** (10 points): This section documents any insights or lessons learned from this Informatics assignment

   g. **Appendix** (10 points):

      i. This section describes your overall approach to implement the Python programs and the list of all Python programs that you implemented and their purpose.

      ii. This section must also include the 3-page proposal.

4. **Data and Programs** (50 points): For each team:

   a. You are required to handin a screen capture of your "testing session" using your programs. (10 points)

   b. You are required to handin all program files. (10 points)

   c. You are required to handin all input and output files. (5 points)

5. **Final Project Demo** (50 points): For each team:

   a. You are required to present your Final Project using Powerpoint slides and execution of your programs at real-time on Demo Day (December 3, 2018)

   b. Your presentation should cover all sections of your Final Project report adequately (35 points)

   c. Your demonstration of your programs should proceed smoothly showing how you obtain the output from running your programs (15 points)

6. You are required to handin online the above using **http://cse.unl.edu/handin/**