Working With Data

CSCI-GA.1121

Fall, 2017: Professor Deena Engel

Course description:

Knowing how to work with data in many formats is an essential skill for humanists and social scientists in their study of the world. In this course, students study the principles of database design and learn to build, populate, manipulate, and query databases based on datasets relevant to their fields of interest, using a project-based learning approach. Students will also explore data presentation through data visualization.

No Prerequisite.

Overview:

In this course, we introduce principles and applications of database design, implementation, and data analysis. We begin by discussing the many data formats that humanists and social scientists may encounter, such as text and CSV files, JSON files, XML, and others. Students study concepts in relational and document-oriented database design (currently using SQLite, MySQL and MongoDB), working with substantive examples across a variety of fields. Students will design an SQL database that is specifically tailored to a project of their own design. Students also gain experience with NoSQL and linked data in this course by building projects with data relevant to their fields of study. In the final unit of the course, students will study a variety of data visualization tools and applications to augment their data analysis.

Note: As programming languages and database paradigms evolve, the particular database technologies taught in this course may change.

Sample Projects

Unit 1: Introduction

Annotated bibliography: Students begin by researching and writing an annotated bibliography on data sources relevant to their fields of interest. Students will be asked to continue to add to this annotated bibliography throughout the semester.

Sample Project: Students will select a data set relevant to their field of study. In this project, students will evaluate and "scrub" the data. Students will also have the opportunity to compare and contrast different data formats available. Sample data sources include the Museum of Modern Art collection (https://github.com/MuseumofModernArt/collection), The Victoria and Albert Museum API (http://www.vam.ac.uk/api), NYC Tree Data and other collections from NYC Open Data (http://opendata.cityofnewyork.us) and others.

Unit 2: Databases

SQL - Sample Projects: Students will be asked to write one or a series of MySQL scripts to create, populate and manipulate data from a source identified in their annotated bibliography. Students will be asked to write a series of queries to interrogate their data and write a report summarizing their findings. These data sets could consist of the holdings in a museum or archival collection; data relevant to a sociological study on eldercare; data based on longitudinal education studies; or other fields.

NoSQL

Sample Projects: Students will work with the same or a different dataset to create, populate and manipulate data from a source identified in their annotated bibliography. Students will be asked to write a series of queries to interrogate their data and write a report summarizing their findings. Topics could include datasets that include geospatial data on the sizes and locations of museums throughout the United States or another country to study the effects of the presence/absence of art on the communities; analysis of readership, sales or other data available about the role of novels in a given culture or society over a specific period of time; the possible correlation between New York City property values and public school test scores over time in specific neighborhoods; or other topics.

Linked Data

Students will explore applications in the humanities and social sciences that use linked data and study the significance of this paradigm in their respective fields.

Unit 3: Data Visualization

Sample Projects: Students will design and implement a data visualization project based on a dataset they explored earlier in the semester. Sample projects could include using GIS software to build an interactive map to visually demonstrate the spread of a disease, the presence of artifacts or cultural monuments or the reach of education funding; using Gephi or a similar software application to build a network diagram of a community of artists, writers or scientists; or build charts and diagrams to visualize other statistical results from their research above.

Format:

The class will meet weekly for 2.5 hours, with 1.25 hours allocated to discussion of class readings and 1.25 hours allocated to project-based work. There will be an additional weekly "virtual office hour" wherein students and the instructor can share screens as well as an in-person collaborative office hour in which students can work in small groups with the instructor or a qualified teaching assistant present for assistance.

Students will be expected to read and annotate texts before class, and to ask and answer questions of and from other students before class, using an online learning platform. Classroom interactions will be facilitated with interactive learning software.

Readings:

Chodorow, Kristina *MongoDB: The Definitive Guide, 2nd Edition* Published by O'Reilly 2013.

Churcher, Clare *Beginning Database Design: From Novice to Professional* Published by Apress, 2007.

Tahaghoghi, Seyed M. and Hugh E. Williams *Learning MySQL* Published by O'Reilly, 2006.

Optional: (Books to be put on reserve in the Courant or Bobst Library)

Dewar, Mike *Getting Started with D3: Creating Data-Driven Documents* Published by O'Reilly Media, 2012.

McCallum, Q. Ethan *Bad Data Handbook* O'Reilly, 2013.

Tufte, Edward *Visual Explanations: Images and Quantities, Evidence and Narrative* Cheshire, CT: Graphics Press, 1997.

Tufte, Edward *Beautiful Evidence* Cheshire, CT: Graphics Press, 2006

Course materials: software

1. MySQL: MySQL will be available to students on a Unix/Linux server. Each student will have an account on this server with appropriate storage available for datasets for study.
2. SQLite will be available to students on a Unix/Linux server. Students can also run SQLite locally on their own computers.
3. MongoDB: MongoDB will be available to students on a Unix/Linux server. Each student will have an account on this server with appropriate storage available for datasets for study.
4. Topics in Data Visualization: The selection of open source software will reflect current trends and available packages at the time the course is taught. Current suggestions include:
    - Gephi (http://gephi.github.io) for network analysis
    - D3 - Data Driven Documents (http://d3js.org)
    - A python module such as bokeh (http://bokeh.pydata.org/en/latest/) for students who have some programming background or are interested in working from these models.
    - *Additional applications may be added, time-permitting*

Updated on April 20, 2017