CS 207 - Data Science and Visualization
Project Details

Prof. Sorelle Friedler

Spring 2016

Requirements

Students will work on a semester-long project to analyze a chosen data set. Throughout the semester, they will apply the analysis techniques they are learning in lab to this data set. At the end of the semester, they will produce a research paper containing the most successful of these analyses and visualizations and at least one new analysis and visualization, a description of their methods and the analysis performed, and an explanation of their findings. This paper will be written individually. They will also collaborate with other classmates working with the same data set to create a poster displaying their visualizations.

Data Sets

Students will choose the data set they would like to analyze from the following list or will provide their own data set in consultation with the professor. These choices must be made within the *first two weeks* of the course. Students interested in the scientific computing concentration should choose the scientific data set.

Data sets:
1. Dark Reactions data set - Materials chemistry experiments and associated experimental results.
   Provided by Phil Adler, Alex Norquist, Josh Schrier, and Sorelle Friedler.
2. Monument lab (`http://www.monumentlab.com/`) - a collection of responses to the question, "What is an appropriate monument for the current city of Philadelphia?"
   Provided by Paul Farber.
3. Walter Benjamin on Lancaster Avenue course data set.
   Provided by Andrew Friedman.
4. Recidivism of prisoners released in 1994 (`http://www.icpsr.umich.edu/icpsrweb/RCMD/studies/3355`).
   Provided by Sorelle Friedler.
5. Honor Code data.
   Provided by Brian Guggenheimer.
6. Challenge: Zapotec / Spanish transcriptions.
   Provided by Brook Lillehaugen.

Deadlines and Details

In addition to the data set analyses and visualizations that will accompany each lab, there are the following project deadlines (no extensions will be given):

- Wednesday, April 27th from 2:30-5pm - Each group working with the same data set will present a poster in Zubrow commons. The posters must be printed *by appointment* in the KINSC office on Monday, 4/25 and Tuesday, 4/26.
- End of the exam period - Each student is individually responsible for handing in a research paper detailing their findings (analyses and visualizations) about the data set they studied. The research paper should be 5 - 10 pages in length (single spaced, 12pt font) and include:
  - an introduction (to the data set and the questions considered)
  - an analysis section (describing the cluster or network analysis techniques used in detail). This section should include illuminating analyses from the lab work during the semester as well as at least one new analysis.
  - a visualization section (describing the visualization techniques used and showing the resulting visualization). This section should include illuminating visualizations from the lab work during the semester as well as at least one new visualization.
  - a results section (summarizing the answers to the questions considered)

  These results and visualizations should include the *good* results from the lab work as well as one *extension* of the lab work specific to the project. They should also incorporate feedback from the poster session.

Grading

The project, which counts for 40% of each student's course grade, will be graded in the following parts:

- Lab work (30%) - this includes the portion of each lab, done individually throughout the semester, which works to analyze / visualize the student's data set.
- Poster (15%) - each student in the poster group will receive the same grade.
- Final paper (55%) - each student will write this paper individually.