SI 618 PROJECT PART 1 REPORT

This document describes how your project-1 reports will be graded. First and foremost, following the individual original work policy clearly stated at the start of the course, the topic and questions you ask in your project must be of your own invention. **If you used ideas from a particular web site or previous project, or did your project as part of an existing research collaboration, you must identify your sources and/or collaborators and provide links and citation(s) where appropriate.**

As a guide, the report should probably be not much more than 3-10 pages depending on space used for any visualizations, tables, etc. I've put three examples from past years on Canvas. The format of the report is flexible in that you can include additional information, but at a minimum it should have the following sections listed below.

1. Motivation (8 points): (a) Briefly state the nature of your project and why you chose it. (b) What specific question or goal did you try to address?
2. Data Sources (16 points): Describe the properties of the two dataset(s) or API services you used. Be specific. Your information at a minimum should include but not be limited to:
    1. Where the datasets or API resources were located,
    2. What formats they returned/used,
    3. What were the important variables contained in them,
    4. How many records you used or retrieved (if using an API), and
    5. What time periods they covered (if there is a time element)

For example, if you downloaded data or used API services, you should state the specific URLs to those files or resources. It should require zero effort on my part to find and access the exact resources you used if I need to do so.

3. Data Manipulation Methods (24 points): For each of your two sources, describe how you manipulated the data. For example:
    1. How specifically did you need to manipulate the data?
    2. How did you handle missing, incomplete, or incorrect data?
    3. How did you perform conversion or processing steps?
    4. What variables and steps did you use to join the two data resources to perform your data analysis?
    5. Briefly describe the workflow of your source code and what the main parts do.

(Some of these might not apply to you, just say it doesn't apply, if so.)

4. Analysis and Visualization (44 points):
    1. A key goal of this project was bringing together two different data resources to answer three interesting questions or find new insights that could not have been answered with either data resource alone. Now describe the analysis steps you performed on your combined dataset to address that goal/question. ***Do this for the three questions separately***. Remember that you needed to use large-scale computation techniques (e.g. mrjob, spark). Be specific, and include references to key functions or parts of your code.
    2. What interesting relationships or insights did you get from your analysis?
    3. What didn't work, and why?
    4. For at least one of the three analyses you worked on, include a visualization (chart, plot, tagcloud, map or other graphic) that summarizes your analysis.

5. What challenges did you encounter and how did you solve them? (8 points): Describe the challenges you encountered and how you solved them.

WHAT TO SUBMIT: Please submit everything used for your project in the usual manner, by including it in a ZIP file project_part1_report_***youruniqname.zip***

Here are the files you need to submit:

- Report for part-1 named *si618-project-part-1-youruniqname.pdf*
- Source code for part-1 under folder *youruniqname-part-1*
- A folder containing data files and/or links to datafiles: For each data file used, include one of the following: 1) the data file itself if it is small, 2) a sample file containing the first 1000 records, or 3) working URLS that point to the data. If your data is sensitive (not publicly available), create a synthetic dataset that can allow us to test your code if necessary. If your data is sensitive and you cannot share, please contact me and confirm. In that case, you can include a readme file in this folder stating that.

As part of the grading the instructor and/or GSI may attempt to reproduce your results using your code and data, and you are expected to assist with this if we request it.