Big Data – Spring 2018
Project Administration, Review and Evaluation

1. Project Mechanics and Team Forming

Projects are performed in teams of three (3) students. We have provided a list of projects you can choose from in https://docs.google.com/spreadsheets/d/1QWSJr29I4mAZEAqqhxyqNJLIXWJZprGBQeIngupjUmc/edit?usp=sharing

Most of these projects can be carried out using open data sets, such as https://opendata.cityofnewyork.us

You must use either Hadoop or Spark and work on a large-scale problem that cannot be solved using a single machine. Your code and scripts must be made available on GitHub and the outputs of your project should be reproducible -- you should include enough information so that others can reproduce what you did.

2. Milestones

March 13: Submit a form with the information for your group at: https://drive.google.com/open?id=1w54Y-icqEXRCenYxgpakKo8Jj-X_zKmTg-mu4ejCnW4

If you want to work on a project not on the list, it must be approved by the instructors. In the form, you need to provide a link to a Google Doc with the description of the proposed project (including the problem statement and data that will be used), and explain how the project meets the stated requirements. Make sure the Google Doc is readable by anybody that has a link to it.

March 19-20: Prepare a 1-page proposal and submit to NYU Classes indicating 1 day before your meeting: your choice for the project, previous work and references, problem description and goal, the data sets you will use, the method/algorithm/task you propose, and evaluation criteria. The evaluation metrics are as follows:

- Previous works and references – 2 points
- Understanding the problem, its formulation and goal – 2 points
- Dataset to be collected, method/algorithm proposed – 2 points
- Evaluation criteria – 2 points
- Writing clarity and structure – 2 points

April 16-17 (day before your meeting): Prepare a 3-page document and submit to NYU Classes describing your preliminary results and any updates. The evaluation metrics for the milestones report are as follows:

- Introduction – 2 points
- Problem formulation – 2 points
- Related works and references – 2 points
- Methods, architecture and design – 4 points
- Preliminary results – 3 points
- Technical depth and innovation – 3 points
- Code repository, correctness, and readability – 4 points

The document must follow the ACM Proceedings Format, using either the **sample-sigconf.tex** template provided at https://www.acm.org/publications/proceedings-template for LaTeX (version 2e).

May 7th and 14th: Final project presentations will be done in a poster format: instructors and your classmates will be your audience. All teammates must be present on the day of their team's presentation. Instructors will make assign the projects on the different days. Plan for a 10-minute presentation, including time for questions. The presentation will count 20 points to your project grade:

- Slides, visualization, appearance of images, legibility (6 points). Tips: limit each slide to 8 lines of text or less, with standard fonts of 28 points or more with large spacing
- Delivery (6 points).
- Content, organization (8 points).

May 14th: Final project report due – submit both the slides and report in NYU Classes (50 points). The final report must follow the ACM Proceedings Format, using the **sample-sigconf.tex** template provided at https://www.acm.org/publications/proceedings-template for LaTeX (version 2e).

The evaluation metrics for the final report are as follows:

- Introduction – 5 points
- Problem formulation – 5 points
- Related works and references – 5 points
- Methods, architecture and design – 10 points
- Results – 10 points
- Technical depth and innovation – 5 points
- Code repository, correctness, and readability – 10 points

Review questions for graders

1. Briefly summarize the project.
2. What are the key strengths or positive aspects of the work?
3. What are the limitations of the work?
4. Does the report consider previous approaches or are there major works missing?
5. Is the problem clearly stated? Does it make sense?
6. Is the project related to big data and the topics covered in class?
7. Is the method clearly described? If not, which paragraphs or statements are unclear.
8. Is there a new or useful component to the project?
9. Is there a Github code repository? Is the code readable and working? Are the results/outputs described in the report reproducible?