Big Data – Fall 2020
Project Administration, Review and Evaluation

1. Project Mechanics and Team Forming

Projects are performed in teams of three (3) students. You must use either Hadoop or Spark and your work/solution must scale for large data. Note that you can do analysis and visualization of the results you produce using Hadoop/Spark results using your laptop/desktop.

Your code/scripts must be made available on GitHub and the outputs of your project must be *reproducible* -- you should include enough information so that others can re-run and reproduce the results you report.

2. Milestones

Nov 13th: Submit the information about your group and your project selection using this Google Form:
https://docs.google.com/forms/d/e/1FAIpQLSenLCWGmxWamJGKVw2am1XrE0zzzJEmD_fsvu_NzNbYkAbPug/viewform?usp=sf_link

You can find a list of projects at
https://docs.google.com/document/d/1rXxbGV7SAvkvzspiN7XY25ukoL6su-8FMjbG8pvf0z0/edit

To help you find a group, I have also created a spreadsheet where you can add your name under the project(s) you are interested in:
https://docs.google.com/spreadsheets/d/1xXsM9fRB96Lo8scAwXGHsJwA0xTbTgoWlwpjEAT6Fn8/edit#gid=0

We will also have a group matching section during class.

Nov 23rd: Prepare a 1-page summary and submit to NYU Classes indicating:

- your choice for the project,
- previous work and references,
- problem description and goal,
- the data sets you will use,

- the method/approach you propose,
- evaluation criteria,
- week-by-week schedule with milestones for the different group members.

You must maintain a file in your git repository named “milestones.txt” that lists your milestones and that is updated weekly -- the file should clearly show the tasks accomplished and tasks that are delayed.

Dec 7th: Project report is due. For the report, you should follow the format of a research paper (see suggested outline below), and I suggest (but do not require) that you use LaTeX (Overleaf) and the ACM format (https://www.overleaf.com/gallery/tagged/acm-official#.WOuOk2e1taQ).

You are not expected to complete a paper that is ready to be published, but I expect your report to be a starting point for a publication.

The suggested structure for the report and evaluation metrics are as follows:

- Introduction – 5 points
- Problem formulation – 10 points
- Related work – 10 points
- Methods, architecture and design – 30 points
- Results –25 points
- References (cited in the report and related work)

In addition, we will evaluate

- Technical depth and innovation – 10 points
- Code repository, correctness, and readability – 10 points

Dec 7th-14th: Project presentations (20 points)
You will give a 10-minute presentation about your project in which you will summarize your key findings/contributions. You should use Google Slides for your presentation and include the names of the group members, project name, and link to github repo.

Here's a suggested outline:

- State the research question(s) you investigated
- Describe the method and data you used to answer the question(s)
- Present your findings and insights you gained
- Discuss challenges you encountered, and limitations of your approach

You should not repeat everything that is in your report. Instead, focus on what you deem is most interesting about your method, findings and the experience you gained.

Your instructor and classmates will be your audience. All teammates must be present on the day of their team's presentation.

Groups will be randomly assigned to one of the presentation dates.

Your presentation will count towards 15% of your project grade.

Report 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?