B669 PROJECT INSTRUCTIONS [*]

1 Overview

The specifics of the project will be very flexible. During the course many problems (in particular, discussed in Lecture 3) will be introduced in various computational models for Big Data. A few of them will be discussed in detail, and the rest will only be mentioned briefly. For the project, you can choose one in the following three directions.

1. (Survey) pick a problem that is only briefly mentioned in the class and make a survey of its state-of-art results.
2. (Implementation) pick some algorithms that are mentioned in the class, implement and compare them (as well as other algorithms that you can think of). No requirement on the programming language; you can use any language you want.
3. (Research) propose new algorithms for problems in models that are discussed in the course. You can either analyze them theoretically (that is, prove some bounds on space/time/communication), or implement them and compare with existing algorithms.

The grade of the projects will depend on how difficult the task is (e.g., proposing good new algorithms will generally be more difficult than understanding/implementing existing ones), and how well it is done.

Your project will consist of the following elements.

1. Project Proposal: **Due 11:59pm, September 22, 2024**
2. Intermediate Report: **Due 11:59pm, October 20, 2024**
3. Final Report: **Due 11:59pm, November 24, 2024**
4. Presentation: **December 6 - December 13, 2024, in class**

[*] B669 SUBLINEAR ALGORITHMS FOR BIG DATA. INSTRUCTOR: QIN ZHANG

2 Project proposal (-5 points if not submitted).

Prepare a 1 page document (use Latex) detailing your plan. This does not need to be too detailed, but needs to contain:

1. Which direction (survey, implementation or research)?

2. What problem you plan to study?

3. Why this problem is interesting?

4. If you choose to do survey, which papers you plan to read?

5. If you choose to do implementation, what data you plan to use and where you plan to get it from?

6. If you choose to do research, what will be new, or what the instructor will learn?

It is quite likely the instructor will provide feedback and alter or modify your proposed plans. This can either happen by students stopping by to discuss with the instructor before the proposal is due, or will come in feedback on the specific proposal. This step is most important when the topic is related to material that is covered later in the class.

If you choose to do a survey, you may want to take a look at the list of papers that has been posted in the course website, and references therein.

If you choose to do some implementation, here is a list of datasets that you can explore.

- `http://snap.stanford.edu/data/`
- `http://www.census.gov/`
- `http://data.geocomm.com/catalog/`
- `http://meta.wikimedia.org/wiki/Data_dumps`
- `http://ngrams.googlelabs.com/datasets`
- `http://kdd.ics.uci.edu/`
- `http://www.cs.utah.edu/~lifeifei/datasets.html`
- `http://www.cise.ufl.edu/research/sparse/matrices/`
- `http://webscope.sandbox.yahoo.com/`
- `http://www.google.com/publicdata/directory`

- `http://www.infochimps.com/tags/twitter`
- `http://lib.stat.cmu.edu/datasets/`

If you have an advisor, they may also be good sources of problems and data.

If you choose to do new research, you need to talk to the instructor when making the proposal.

3 Intermediate Report (20 points)

Prepare a 1 page report (use Latex) describing your progress so far towards your proposed goal.

Basically this intermediate report should demonstrate to the instructor that you have made non-trivial progress towards your goal.

If you choose to do a survey, you should report what papers you have read and plan to include into the survey. Perhaps you have already set up the structure of the survey, and started working on the details.

If you choose to do some implementation, please describe what data you have collected or are continuing to collect. Please report:

1. How you obtained your data?
2. How large is your data?
3. In what format are you storing your data (be precise)?
4. Did you need to process the original data to get it into an easier, more compressed format?
5. How would you simulate similar data?

In many cases you will not store the data in the original format. That is, you will need to process it to be in some abstract representation (e.g., a matrix, a graph, or a point set). This is usually the most challenging aspect for students. These decisions should be discussed to answer steps 2 and 3 (above) fully. Step 5 is to make you think about how you would model your data. The structure you hope to find is likely correlated with how you model your data. This will also be important if you want to generate synthetic data to see how your technique scales beyond the real data set you have gathered.

You should take a basic algorithm and then modify/improve it. Some basic plots or numbers from experiments that you ran (mainly to convince yourself) that shows everything is working can be included. Perhaps you have finished all of the coding and setup already and just need to run experiments at this point. Then note that and discuss what suite of experiments you plan to run for the final report.

If you choose to do new research, you need to talk to the instructor.

The instructor will attempt to provide feedback to make the final reports as strong as possible. Thus the more progress you have made and the more information you include, the more success you will likely have with your final report.

4 Final Report (30 points)

If you choose to do a survey, there is no page limit. Remember, at any time, you **should not simply copy** materials from the papers you survey (otherwise it will be considered as plagiarism). Usually, you want to summarize/simplify/synthesize algorithms/analysis/results from those papers.

If you choose to do some implementation, your report should be 4 pages, single columned at 11 point or larger font. However, you will be allowed an unlimited number of pages for references and appendices. The report will be graded on the first 4 pages, but additional information to support the first 4 pages, may be appended and referred to. The instructor will only read the appendix at his discretion.

If you choose to do new research, you need to talk to the instructor.

Contents of the report

1. Explain the problem and motivation. If you prepared a thorough proposal and intermediate report, then you may be able to borrow some material from there.

2. If you choose to do a survey, explain what papers/algorithms you have covered. What is the start-of-art results on the problem you study. Can we get better methods/algorithms? Please give your conjecture.

3. If you choose to do some implementation, explain what data you explored? Where did it come from, how did you process it? If you simulated to scale the experiments, how did this work? If your data collection report was thorough, you can likely reuse much of this material.

4. If you choose to do new research, explain what you did. Did you prove something? Did you extend something? What is the key new idea your project is built upon? What are your results?

5. Explain what you learned.

5 Presentation (20 points)

You should make slides and give a talk in front of the class. I hope to see the following elements in your talk.

1. What is the problem and/or data you worked on. Or, what papers you have surveyed?
2. What were the key ideas in your approach?
3. What techniques did you use?
4. What conclusions you came up with? What did you learn?

This is a great opportunity for the class to learn about a large variety of topics. If you approach this presentation as a teaching experience, you will be more likely to succeed.

The format of the presentation is 30-mins talk plus 5-mins Q&A.