B490 PROJECT INSTRUCTIONS *

Spring 2014

1 Overview

The specifics of the project will be very flexible. During the course we will discuss many problems and algorithms (solutions), e.g., locality sensitive hashing, streaming algorithms, clustering algorithms. In assignments you will be asked to implement (if you don’t want to simulate by hand) and run these algorithms in small data. In the project, you can pick one (or more) of them and run it (them) in the real data sets. It could be much fun.

The grade of the projects will depend on how difficult the task is, and how well it is done.

Your project will consist of the following elements. The due dates will be published on the course website.

1. Project Proposal
2. Data Collection Report
3. Intermediate Report
4. Final Report
5. Presentation

Note that some topics will not be covered before many elements of the project are due. This could be an issue. However, typically, most work on a project is crammed in the last week or two of the semester, leading to weak projects. To assist you to choose a topic, the instructor will give a high level introduction of all topics that will be covered in the course. You are highly welcome to choose a topic which will be discussed in the late stage of the course, and the instructor will point to you relevant materials so that you can study them early.

You can work by yourself or in a group of at most 3. If you work in a group of 2 or 3, then the work has to be doubled or tripled.
2 Project proposal (10%)

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

1. Who are in your group (yourself or a group of 2/3)?
2. What problem you plan to study?
3. Why this problem is interesting?
4. What data you plan to use and where you plan to get it from?

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.

Here is a list of datasets that you can explore.

- [http://snap.stanford.edu/data/](http://snap.stanford.edu/data/)
- [http://www.census.gov/](http://www.census.gov/)
- [http://ngrams.googlelabs.com/datasets](http://ngrams.googlelabs.com/datasets)
- [http://www.cs.utah.edu/~lifeifei/datasets.html](http://www.cs.utah.edu/~lifeifei/datasets.html)
- [http://www.cise.ufl.edu/research/sparse/matrices/](http://www.cise.ufl.edu/research/sparse/matrices/)
- [http://www.google.com/publicdata/directory](http://www.google.com/publicdata/directory)
- [http://lib.stat.cmu.edu/datasets/](http://lib.stat.cmu.edu/datasets/)

If you have an advisor, they may also be good sources of problems and data.
3 Data Collection Report (10%)

Prepare a 1 page report describing 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.

4 Intermediate Report (10%)

Prepare a 1 page report describing your progress so far towards your proposed goal. If you work in a group of 2 or 3, the report should be 2 or 3 pages.

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

Perhaps you proposed to take a basic algorithm and then modify/improve it. Then this intermediate report should note that you have gotten the basic technique working. 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.

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.
5 Final Report (40%)

Your report should be 4 pages (7 pages for a group of 2, 10 pages for a group of 3), 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 primarily be graded on the first 4 pages (first 7 pages for a group of 2, first 10 pages for a group of 3), but additional information to support the first 4 pages (first 7 pages for a group of 2, first 10 pages for a group of 3) may be appended and referred to.

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. 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.

3. Explain what you did? In particular, your implementation and the experimental results you obtained. You should present the ideas of your implementation in words instead of cut-paste your codes. Your results should be organized in a table.

4. Explain what you learned.

6 Presentation (30%)

You should make some slides and give a talk in front of the class. If you work in groups, then each of your group members need to participate in the presentation (the amount of time given will be doubled or tripled). I hope to see the following elements in your talk.

1. What is the problem and data you worked on?

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.