Project*

Final Report Due: Monday, December 5, 2011
Turn in report at the start of class.

1 Overview

Your project will consist of three elements.

- Project Proposal : Due October 31, 2011
- Project Report : Due December 5, 2011
- Project Presentation : Either December 5, 7, or 9, 2011.

As in any research in order to get people to pay attention, you will need to be able to present your work efficiently in written and oral form.

You may work in teams of up to 3, but the amount of work you perform will need to scale accordingly. All students will need to have clearly defined roles as demonstrated in the final report and presentation.

1.1 Scale of Project

The specifics of the project will be very flexible. I expect each student to explore some topic in this class and to demonstrate or extend techniques from this class within that topic. Alternatively, student may explore a specific project using multiple large data paradigms. Projects may be experimental or analytical, as long as it is clear the student has demonstrated effort in exploring a topic.

2 Project Proposal

Due Wednesday, October 31, 2011

Prepare an at most **1 page** document detailing your plan. It can be less than 1 page, usually half a page will suffice.

If you plan to attack a specific problem in a large data paradigm, (1) describe the problem you plan to attack, and (2) why you think using that paradigm will be effective. I expect either this to have not to have been done before (in published form), or if it has been done, explain how you hope to extend this technique, or what unexplored aspect of the project you hope to explore.

If you plan to compare a certain problem using multiple paradigms, (1) describe the specific problems, (2) if it has been done within any of those paradigms before, (3) and which advantages you expect from each paradigm. (You may discover you expectations are wrong, but I want you to think about it ahead of time.)

If you plan for more analysis, (1) describe the specific problem, (2) what sorts of analysis bounds you expect to find. (Again, these can be wrong - thats research - but I want you to think through whether the answer is obvious).

The project can be related to or part of a larger ongoing research project. In this case, I also expect the proposal to describe (1) what has already been done in the larger project, (2) how this fits into the larger picture, and (3) what is new and different about the class project.

2.1 Timing

The instructor reserves the right to return any proposal, with explanation, and ask the student to create a new one. This can happen if the instructor finds the scope of the project way too large, or way too small. Or it can happen if a proposal too similar has already been submitted (first come, first serve).

For this reason, **it is highly recommended that turn in the proposal well-before the deadline**. If you submit the proposal for the first time at the deadline, and it is returned, you will be at a distinct disadvantage.

2.2 Topics

The best projects occur on topics that students have experience with and/or are passionate about. First look within your own experience.

If you still have trouble finding a topic I suggest exploring these resources for interesting datasets to explore. You cannot just explore a data set, you must also propose to solve some algorithmic problem on that data.

- `http://snap.stanford.edu/proj/socmedia-kdd/`
- `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/`

3 Project Report

Due Monday, December 5, 2011

Your report will 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 four pages, but additional information to support the first four pages may be appended and referred to. The instructor will only read the appendix at his discretion.

If you work in a group of more than 1 student, then your report will be 4 pages per student, plus unlimited space for references and appendices. It should be painfully clear what part of the project was worked on by each student. If the distribution of work was complicated, a table in the appendix may be a good idea.

Why only 4 pages? A key aspect of scientific writing is efficiently conveying information. I expect students to easily generate more than 4 pages of information, but you need to convey this information to me efficiently. What are the key ideas? What are the key experiments to show me plots for? What is the relevant related work to highlight?

3.1 Content

1. Explain the problem and motivation. If you prepared a thorough proposal, then you may be able to borrow some material from here.
2. What is the key idea your project is built upon? If there is no interesting ideas, I will be a little disappointed.
3. Explain what you did. Did you prove something? Did you implement something? Did you compare several things? Did you extend something?

4. Explain what you learned. This can be expressed as charts of experiments, or proofs of a theorem. But you should also include what lessons you came away with in words; just charts or mathematics is insufficient.

4 Project Presentation

Given on Monday, December 5; Wednesday, December 7; or Friday, December 9.

You will have **7 minutes per student** to present what you did. I expect each student to present for 7 minutes (a group of 3 can not have 1 person present for 21 minutes). We will have 3 classes of 80 minutes each for about 30 presentations. So we will have 1 minute for questions and transitions between presentations. So if things go long, the instructor will, unfortunately, be forced to cut students off.

As with writing, efficient presentations is a very important part of research progress. Students will have to efficiently and effectively present their project.

I hope that the three days of presentation will be some of the most exciting classes. I expect to hear about a lot of different large data sets and algorithms for dealing with those data sets. The combined experience of the class has the potential to provide great insight into the array of techniques and algorithms for dealing with massive data.

4.1 Content

I expect to hear three things in your presentation:

1. What is the problem and data you worked on?
2. What were the key ideas in your approach?
3. 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.

4.2 Schedule

There are three presentation days:

A : Monday, December 5
B : Wednesday, December 7
C : Friday, December 9

A student will have a chance to sign up for day A as soon as their proposals are approved. When 10 students have signed up for day A, then students who have had a proposal approved will be able to sign up for day B. The remainder of students will be assigned day C if 20 students have signed up for day A or B.

If by the start of day A, when the reports are due, fewer than 10 people have signed up for day A or B, the remainder of assigned slots will be chosen at random.

At the start of day A, the order for all three days will be assigned arbitrarily.