Data Mining - Project Guideline (Subject to Update)

All project should have:

- Data. This is the most important part of the project. Data should have reasonable size to be useful, say at least 1000 (1k) data points. Though the threshold 1000 is somewhat ridiculous, I find it hard to have an interesting idea with small data set. If you are planning a project with a data set less than 1000 points, you should reach out to me to make sure your ideas with the data are interesting.
- Problem. You should be very clear about the problem you want to solve given you data set. The problem should not be the same as the ones in the programming assignments, but you could use the method (hence code) in the programming assignment to help tackle your interesting (and bigger) problem.

Expectation for SENG 474 The general idea is you have a good dataset, you have a good problem. You're able to apply some techniques (may be learned in class, may be not) to the data set, and drawn some interesting conclusion. Some questions that may help you to formulate a good problem is:

- What aspect of the data set are you interested in knowing?
- Why do you choose a particular method, or an algorithmic technique?
- What is your original assumption about the data, and up on mining the data, how is final result different from what you're assuming. That is, does the final result confirm your belief about the data? If it does not, then why?

Note that the above questions just serve as an illustration. You may (and I encourage you to) come up with far more interesting questions.

Expectation for CSC 578D In addition to what are expected for a SENG 474 project, I expect that CSCS 578D projects are bigger and deeper. Up on having data and problem, proposing a solution for it, you should look more into the results, try to understand it, present your findings and based on the findings, ask what else can be done to better solve the problem.

This guideline is very rough, and I do expect that you may not follow it, because I do not want to constrain your creativity in these rules. As long as you think what you're doing is surely interesting, it should be ok. You're welcome to talk to me about your potential ideas for projects to get feedback.

1 Project Report

Every group is expected to write a report for the project. You should submit the pdf version of the report to Connex, along with the code of the projects. Please **be precise**. The maximum length of the main content report is 3-page. By main content, I mean excluding the reference section. If you think you need more than 3 pages, let me know, with a reasonable explanation.

Here would be a couple of things that I am looking for from a report:

- Which problem are you trying to solve why?
- Where and how do you obtain the data? How big is your data?
- What are your ideas to solve the problem?
- What is your hypothesis for the ideas to work? A more interesting question is how do you verify your hypothesis?
- How does the result look like? Does it confirm your hypothesis?
- What have you done to make your original ideas better?
- What is the running time of your algorithm? Is your algorithm scalable?
- If you are given more time, what can be done to even improve it further?
- What have you learned from the project?

Of course your report does not need to address all of the above questions, but it should address as many as possible. Also, there could be other aspects of your project that do not fit into any question above and I would be very interested in learning about that.

Also, please acknowledge relevant materials in your report. There is *no limit* on the length of the reference section. For example, you report can be up to 10 pages, with 7 pages of reference.

2 Deadline

My idea is to have all of you present your projects during the last week of the class. I am working toward scheduling the presentation. The presentation time for each group is about 5-7 minutes for each group, so please be concise.

After presentation, you can still working on your projects to improve/finish it. But major ideas and results should be ready by the time of the presentation. The deadline for final submission of your project code and report is **Friday, April 05, 11:55 pm**.