CS425 Course Project

Important Dates:
October 16: Send the TA and tutor the following information: 1) Group members, 2) Three project topics in order of your preference. For each topic, include a few sentences describing the project topic, which datasets you are going to use, which programming language you will use, etc.
Dec 4-Dec 22: Project presentations
Dec 21: Final reports due
Dec 21-Dec 23: Project demos

Contact Information:
TA: H. Gokhan Akcay (akcay@cs.bilkent.edu.tr)
Tutor: Orcun Gumus (orcun.gumus@epfl.ch)
Office hours: Send an email to both the TA and tutor to set up appointments.

Project Overview:
You can choose either 1) an implementation project, or 2) a survey project, 3) or a small research project. The detailed descriptions for each project type are below. If in doubt, you're recommended to choose an implementation project, which is the safer option for most people. Options #2 and #3 are for students who want to do an in-depth study on a particular topic to see if they want to work on it in the future.

This will be a group project. Each group is supposed to have 3 or 4 people. In your report, you're supposed to explain how each person contributed to the project. Also, 10% of your project grade will be coming from your group members.

We will have project presentations in the last 3 weeks of classes. If you choose a topic that is taught earlier in the semester (e.g. PageRank or duplicate detection), your presentation is expected to be scheduled earlier than others. If you choose a topic that is taught late in the semester (e.g. Apriori algorithm), you are supposed to study the topic yourself so that you can start your project early. This will be taken into account while grading.

All project reports are supposed to be turned in by Dec 21. Also, project demos will be scheduled in the last week of classes. Your midterm plus project grade should be above a threshold for you to be eligible to take the final exam.

Please send 3 project topics to the TA and tutor by Oct 16. For each topic, describe in a few sentences what your plans are, including what type of datasets you are planning to use. We will try to accommodate your preferences, and send you the project topic we choose for you. The purpose here is to balance the number of projects per topic. If you cannot form a group, please contact the TA as soon as possible.

You are encouraged to propose your own topic if you have a different project in mind. You can set up an appointment with the TA and tutor to discuss your project proposals before Oct 16.

Your project grade will be based on: 1) the quality of your presentation, 2) the success of your demo, and 3) quality of your final report, 4) the total effort you put into the project.

Project types:

Implementation project: You can choose to implement an algorithm that is within the scope of this course. This can be one of the example topics below or you can propose an algorithm related to this course. You can choose a dataset that is already available (see below) or you can create a dataset yourself by pulling data from public websites. If you choose to create your own dataset, this will be counted as part of the implementation effort. If you choose an implementation project, it is expected that the algorithm is complex enough and your implementation effort is significant. (e.g. You cannot just reuse someone else's code.) You're expected to show how much effort you put into the project in your presentation and report. In the demo, you are expected to describe your code in detail and show that it works.

Survey projects: Survey papers must be of high quality (similar to the published survey papers in journals). You're expected to study different algorithms on the chosen topic, give an overview of the problem and solutions, explain 3-5 algorithms in detail *with your own words*, and present the open research problems on that topic. During the project "demo", you will be asked about the algorithms in your report, and you're expected to explain them in detail. The report should not contain sentences/figures/pseudo-codes from other papers as is. **You should check the university policy on plagiarism before deciding to choose a survey project.** In your presentation, you're expected to "teach" your findings to the rest of the class. Your report will be the survey paper you write.

Mini research project: If you have an idea in mind and you want to try it, you can choose a research project. You're still expected to do a short survey (not as detailed as in a survey-only project). You can implement your idea on top of an existing implementation (with explicit citation of course), and analyze your results. In your presentation, you're expected to explain the problem(s) you're trying to solve, your proposed solution, and your initial results. Your report should clearly describe the novelty of your idea and the results you obtain with respect to state of the art.

Potential project topics:

1) Implementation: Find scientific articles with most similar citations
2) Implementation: Find documents with most similar contents
3) Implementation: Map-reduce for 3 applications: 1) Matrix-vector multiplication, 2) Matrix-matrix multiplication, 3) PageRank. You are expected to install a virtual machine to your laptop with Map-reduce support and learn how to use the framework. Some pointers will be provided to you to help you get started. However, you will be responsible to make it work on your laptop.
4) Implementation: 3-types of collaborative filtering algorithms: 1) user-user, 2) item-item, 3) latent-factor
5) Implementation: PCY extension of Apriori algorithm
6) Implementation: Study the GraphChi paper, and implement PageRank using GraphChi layout.
7) Implementation: Community detection algorithm on social network datasets
8) Implementation: Community detection algorithm on protein interaction datasets
9) Implementation: Multi-core or GPU implementation of a non-trivial algorithm related to the class.
10) Survey: Association rule mining algorithms.
11) Survey: Distributed graph frameworks such as Pregel, Giraph, GraphLab, etc.
12) Survey: Different collaborative filtering algorithms used in the Netflix challenge
13) Survey: Multi-core and GPU implementations of a non-trivial algorithm related to the class.
14) Survey: Community detection algorithms

If you have other project topics in mind, you should talk to the TA and tutor before the Oct 16 deadline.

Data sets:

It is your responsibility to find or create datasets for your project. This will also count as part of the effort you put into the project. To help you get started, here are some links you can explore to choose a dataset:

https://snap.stanford.edu/data/
http://web.stanford.edu/class/cs341/data.html
https://github.com/caesar0301/awesome-public-datasets
https://www.kaggle.com/competitions
https://inclass.kaggle.com/c/movie
http://interactome.dfci.harvard.edu/S_cerevisiae/index.php?page=download

You can also set up an appointment with the course tutor if you have difficulty finding datasets.