Algorithms for Big Data
Project

Post questions to the project channel of teams

2022

Important

The link for the projects approved is here.

When the paper you have chosen is approved (see below), your name(s) will be added to that spreadsheet by me. You do not edit the spreadsheet directly.

If your name is not there, do not have an approved project topic!

This spreadsheet will also be used to schedule presentations as June approaches.

Project guidelines

Your grade is 50% final written exam and 50% project.

The goal of the project is for you to learn an algorithm not covered in the class, implement it and understand its analysis.

Projects may be completed alone or in a team of at most two people. Both members of the team will get an identical score.

The main deliverables are the code and a 20-minute-long presentation.

The project is due on May 30 at 11:59PM. You will lose 1% per hour of lateness.

The algorithm

You must choose an paper that is

- Relevant to the field of big data.

- Published in a reputable conference or journal in the past 10 years. Examples of reputable conferences where algorithms results appear are can be found here, and examples of non-reputable publishers can be found here Note that in theoretical computer science, conference publication is typically the main way one gains recognition for a result. Recent papers are also posted on arXiv for convenience, but arXiv has almost no editorial filters. After appearing at a conference a paper may also appear in a journal, but the absence of this step says nothing about the paper's quality.
- The algorithm must have a rigorous analysis.
- You have freedom of choice, you can pick something more general or some algorithm developed for a particular domain where big data is prevalent, e.g. bioinformatics.
- I should not be an author of the paper you have chosen.
- No duplicates are allowed. Each group should broadcast their choice on the Project teams thread. Whoever announces first has the right to submit the project on that paper. Duplicate projects who did not announce on teams will get a zero score.
- Don't choose something that has publicly available code.
- Each group must choose a different paper, and you can not choose something that was chosen last year.

Important: You must get approval from me for your paper. Send me a message on teams with the paper you wish to present (or a link if there is a publicly-available PDF). Once approved I will add it to the spreadsheet.

The implementation

You should understand and implement the algorithm. You should test it to see that it works and that the theoretical claims about the algorithm (runtime, space) are true. You should try to compare it against other solutions, either publicly available ones, or "trivial" ones as appropriate.

You code should be in any widely used language.

You should convince me that the code is correct. This could be done graphically.

The presentation

If you do the project in a pair, you must both present at the same time.

You should give with your partner a 20 minute long presentation. In this presentation, you should explain the problem, indicate the model (e.g. streaming, cache-oblivious, etc), the algorithm, what is the theoretical performance of the algorithm, what was known before this algorithm. You should usually show how the algorithm works on an example. The goal is that after your presentation, I should understand the algorithm and its analysis.

You should then present the results of your implementation (5 mins max).

The presentations will be scheduled during the June exam session (May 30-June 24). Note that the university's exam scheduling system is not capable of scheduling presentations done in pairs, there will be a sign-up google doc to choose your time.

To hand in

- Email to bigdata21@johniacono.com
- Subject: "Project submission of XXX (and XXX)"
- CC your partner, if a pair
- Attach a pdf of the paper
- Submit all files needed to run you code, including any data sets you may be using. If not using a notebook like Jupyter, include a README and a Makefile, where the README clearly says what I need to do to run your code as well as any other needed explanations and screenshots of your code running. If files are to large to attach, send a link.
- I will confirm submission via email if you don't get a confirmation in a few hours during the working day, contact me via teams.

Questions

I will, of course ask you questions after your presentation. Allow 40 minutes total.

Grading

Grading is 50% implementation (including the part of the presentation where you discuss your implementation) and 50% presentation (excluding the implementation discussion).

The implementation score will be based on

- The difficulty of the implementation.

- Note that the difficulty will be adjusted based on whether the project is done by one or two people.
- The quality of the code, the documentation
- How well the code implements the algorithms in terms of correctness and metrics such as speed and time

The presentation score will be based on how well you understand the algorithm, its analysis and its historical context, and how well you can communicate these things. Difficulty will also play a role in the scoring.

Second session

The project for the second session is due August 15th at 11:59PM. Presentations will be scheduled during the exam session. If you completed the project for the first session and did not pass the course, you can either re-do the project or keep your project grade from the first session. If you re-do the project you need to pick a new paper.

Academic honesty

All work should be your own or referenced. Reference anything that is not yours. If you copy more than a single line of code from somewhere else include a comment with a URL. If you use an figure in your presentation that you did not make, it must be referenced. This does not mean that copying code is forbidden for some subroutine incidental to your project, it is not! Just give recognition where it is due. Any breach of academic honestly will result in a zero grade for the project.