ENGG5108/CSCI 5510 Project Specification

Project grouping deadline: 23:59:59, Sep 30, 2015
Proposal deadline: 23:59:59, Oct 23, 2015
Proposal peer review deadline: 23:59:59, Nov 6, 2015
Project Presentation time: Nov 25, 2015
Final report, presentation file, source code deadline: 23:59:59, Dec 7, 2015
Submit to engg5108@cse.cuhk.edu.hk

Introduction:

The course project is to give the students hands-on experience on big data analytics. The project is open-ended and you can pick any topic that is relevant to big data analytics. The project thus emphasizes applied research and "deliverables", meaning that the outcome of your project should be something tangible, i.e., typically some kind of prototype system that can be demonstrated or some interesting results obtained from your own analysis of massive data with some design and implementation. In order to choose a suitable topic, you should investigate the state-of-the-art algorithms and technologies via paper survey. Finally, you should demonstrate your project via presentation and finish the final report. Up to 4 students could form a team to finish the project. Please send your group's student names, student ids, and student emails to course account before above mentioned deadline. Tutors would randomly group students if they fail to send group information to the course account before the deadline. We plan to make each team size as close to 4 per team as possible.

Grading criteria:

Your project will be graded primarily based on the following weighting scheme:

- Project proposal and final report: 50%
  - Proposal: 5%
  - Proposal review: 5%
  - Final report: 40%
- Presentation and demo: 50%

Late submissions within three days will be deduced 30% of the score; late submissions more than three days will get 0 marks on that phase.
The factors to be considered in grading include (1) the originality and novelty of the techniques used to solve the problem; (2) the relevance to the course; (3) the challenges you have to solve (i.e., technical contributions); and (4) the quality of presentation/writing. Detailed guidelines about what to include are explained below.

Project requirement:

Since our course manly focuses on analyzing massive datasets and draw useful conclusions from the datasets, the project intends to familiarize you with modern techniques and tool chains in analyzing big data. You will be provided with a budget to create and maintain your own account to access the major cloud computing environment (i.e., Amazon EC2), which is equipped with Hadoop distributed computing facility or other big machine learning systems to carry out large scale computing.

So in your project, you are required to make good use of these big data analytics facilities. To be more specific, you are required to choose one from two different domains as follows:

Application domain:

- You should make good use of the distributed computing environment Hadoop to implement your project.
- The datasets you use should be relatively "large". Nowadays, free academic datasets are generally associated with at least tens of thousands of users in the Internet. So you can find large datasets easily online. We do not have a hard requirement of what size in terms of GB is large since the format of the data also counts. In your project proposal, the details of the dataset should be included for assessment.
- The topics of the project your plan to work on should be related to the topics taught in this course. Some suggested topics include: graph analysis (Twitter, Facebook graph analysis), recommender systems, online clustering algorithms, etc. Take graph analysis as an example, you can utilize community detection skills to analyze user relationship in Twitter.
- You can use existing tools to help your building up your system and facilitate your analysis. You should state clearly which parts of your project are done by existing tools and which parts are your own original work and implementation. Some useful tools are listed in the Tools part.
- Big data have the property of 4Vs: volume, velocity, variety, and veracity. At least one of them should be illustrated in your project. For example, to "volume" property your data may be very large, say, more than 10GB; or your project can address a stream data effectively with respect to the property of "velocity".

Algorithm domain:

- As sophisticated machine learning algorithms play a vital role in data analytics, you are also encouraged to explore the implementation of classical machine learning algorithms over a variety of distributed machine learning systems.
- In this topic, you need to implement a classical machine learning algorithm on several distributed machine learning systems and compare the difference of performance among these systems, which include but not limited to Apache Spark, GraphLab, Apache REEF, Petuum, Pregel+ (see References for details).
- You can choose one of the listed machine learning algorithms to implement:
  - AdaBoost
  - Hidden Markov Models
  - Distributed Sketching
  - Non-negative Matrix Factorization
  - Tensor Decomposition
  - Density-based Spatial Clustering of applications with Noise (DBSCAN)
- If you want to implement other algorithms, please ask the tutors for permission first.
- For algorithm domain topics, you are allowed to only use big synthetic data to test the correctness and the performance of your implementation.

When picking a topic, try to ask yourself the following questions:

- What is the main function or analysis target that you would like to develop?
- What is the motivation for doing your project? Who will benefit from your project and how?
- Is there any existing work that solves the same or similar problem? How do your results compare with others? What are the relative advantages of your project?
- What is the essential goal to be achieved during this project?

Write a proposal:

You are required to write a proposal before you actually go in depth on a topic. Please send your group's proposal to the course account by the due date specified on the above. Each group only needs to send one proposal to the course account. In the proposal, you should address the following questions and include the names, student ids and email addresses of the group.

- What is the motivation of your project?
- Which part of the course topics is most related to your project?
- What are the deliverables you plan to submit by the end of this project?
- Which tentative dataset(s) do you use in your project? Show the basic statistics of the dataset(s).
- What techniques/algorithms will you apply?
- How will you demonstrate the main functions or analysis results?
- What is the existing work that is related to your project? Give a short summary of existing work and state the difference of your project.
- A very rough timeline to show your project milestone. (The timeline does not have to be accurate.)

Review proposals:

Each group will be asked to review several proposals from other groups. You are required to give an integer rating (1-5) for each of following aspects, and also write a detailed comment for each of following aspect. Your proposal review grades will be determined according to your review quality.

- Novelty: Whether the proposed search engine/tool is novel? How is the proposed project different from existing work?
- Significance: Is the problem to be solved important?
- Related work: Are related work thoroughly investigated? Are references of related work provided?
- Technical quality: Do the authors have a plan about what techniques or resources they will use?
- Clarity: How is the overall organization and clarity of the proposal?

Proposal review form will be provided on the course website.

Work on the project:

You can reuse any existing tools or software packages. There are also many tools available on the Internet. See References for some useful pointers. Consider documenting your work regularly. This way, you will already have a lot of materials written down by the end of the semester. Note, however, that you should clearly state the software codes you have adopted from others, and those you have created on your own.

Present the course project:

On Nov. 25, 2015, each project team is expected to make an oral presentation of the project. The purpose of this presentation is: (1) Demo your deliverables and/or present qualitative results. (2) Give you some opportunity to practice presentation skills, which are very important for a successful career. (3) Obtain feedback from lecturers and the tutors of your project. Each group is required to make Powerpoint presentation slides. You are strongly encouraged to implement a demo of your project if possible. Or at least you should show some sample results (e.g., tables, figures, videos). In general, the structure of your presentation should roughly follow your final report.

Your presentation will be graded mainly based on (1) The novelty and significance of your project, (2) The quality of your system and/or evaluation results, (3) The clarity of your slides and presentation, and (4) Whether your presentation has covered all the issues and questions listed in the project requirement. Think about how you can best present your work so as to make it as easy as possible for your audience to understand your main contribution in the project. Try to be concise and to the point. Pictures, illustrations, and examples are generally more effective than text for explaining your project.

Write a final report:

You should write your report as if you were writing a conference paper. You should address the same questions as those you have addressed in the proposal, only with more details, especially regarding some of the challenges that you have solved in developing the search engine/tool. You should also include your system structure diagram or concise screenshots if applicable, and any succinct evaluation results. Furthermore, it would be good to include a brief discussion of how your investigated system can be further extended. You are required to use ACM SIG Proceedings Template with Option 2 (Tighter alternative style) at this link:
http://www.acm.org/sigs/publications/proceedings-templates
You could use either latex or word to write the final report. But your group needs to submit a PDF file to the course account. The page requirement is 8-10 pages. If you have not written such a report before, you may want to take a look at research papers from conferences or journals provided in the reference.
Grading of a project report will be based on following factors:

- The quality of your project as reflected in the importance of problem being addressed, the quality of solution, and the impact of your project.
- The novelty of your project.
- The significance of your project.
- Amount of work that you have done.
- Clarity and completeness of the report itself (i.e., whether you have clearly described what you have done and addressed all the questions that you are suggested to address).

References:

http://www.sigir.org/

http://www2013.wwwconference.org/

http://www.aaai.org/Conferences/conferences.php

http://www.cikm2013.org/

http://www.kdd.org/

http://wsdm2013.org/

http://www.acl2013.org/

http://graphlab.org/graphlab-workshop-2013/

http://recsys.acm.org/recsys13/

http://www.ieeebigdata.org/2013/

http://www.ischool.drexel.edu/bigdata/bigdata2013/

http://icdm2013.rutgers.edu/

Tools:

Apache Spark: Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala, Python and R. In contrast to Hadoop's two-stage disk-based MapReduce paradigm, Spark's multi-stage in-memory primitives provides performance up to 100 faster for certain applications. http://spark.apache.org/

GraphLab: GraphLab is a graph-based, high performance, distributed computation framework written in C++. It contains toolkits covering a variety of aspects: collaborative filtering, clustering, computer vision, graphical models, topic modeling, and linear solvers. http://graphlab.org/

Apache REEF: REEF is a Big Data system that makes it easy to implement scalable, fault-tolerant runtime environments for a range of data processing models (e.g., graph processing and machine learning) on top of resource managers such as Apache YARN and Mesos. https://reef.incubator.apache.org/

Petuum: Petuum is a distributed machine learning framework. It aims to provide a generic algorithmic and systems interface to large scale machine learning, and takes care of difficult systems "plumbing work" and algorithmic acceleration, while simplifying the distributed implementation of ML programs - allowing you to focus on model perfection and Big Data Analytics. http://petuum.github.io/

Pregel+: Pregel+ is not just another open-source Pregel implementation, but a substantially improved distributed graph computing system with effective message reduction. Compared with existing Pregel-like systems, Pregel+ provides simpler programming interface and yet achieves higher computational efficiency. http://www.cse.cuhk.edu.hk/pregelplus/

Scikit-learn: Scikit-learn integrates machine learning algorithms in the tightly-knit scientific Python world, building upon numpy, scipy, and matplotlib. http://scikit-learn.org/stable/

Moa: Moa is a massive online analysis tool for real time streaming data. http://moa.cms.waikato.ac.nz/

Others:

Online learning:

http://hunch.net/~vw/

http://www.cs.huji.ac.il/~shais/code/index.html

http://leon.bottou.org/projects/sgd

Data stream software:

http://people.cs.umass.edu/~dstubbs/streaming/software