CMPT 733 - G100 Programming for Big Data 2

Instructor(s): Steven Bergner, Jiannan Wang

Calendar Objective/Description:
This course is one of two lab courses that are part of the Professional Masters Program in Big Data in the School of Computing Science. This lab course aims to provide students with the hands-on experience needed for a successful career in Big Data in the information technology industry. Many of the assignments will be completed on massive publically available data sets giving them appropriate experience with cloud computing and the algorithms and software tools needed to master programming for Big Data. Over 13 weeks of lab work and 12 hours per week of lab time, and building on the previous lab course CMPT 731, the students will obtain a solid background in programming for Big Data.

Instructor’s Objectives:
From CMPT 726 and CMPT 732, students have learnt machine learning algorithms and big data programming tools. However, when facing a real-world data problem, the students will find that there is still a gap between what they have learnt in class and what they are going to do in practice. The goal of this course is to fill this gap, making the students be able to apply what they have learnt to solve real-world problems. To achieve this goal, our course will cover a set of important topics that a data scientist should know, and teach students about the state-of-the-art approaches. After taking this course, students should feel confident when being asked to extract value from real-world datasets, and know how to ask interesting questions about data, how to choose proper tools, how to design data-processing pipelines, and how to present final data products.

Prerequisites:
CMPT 732: Programming for Big Data 1

Topics:
- Data Preparation
- Data Analytics
- Applied Statistics
- Practical Machine Learning
- Visualization and Communication
- Active Learning and Crowdsourcing
- Deep Learning
- Large-scale Machine Learning

Grading:
Each one of the 8 assignments will count for 9% of the final grade. The final project will count for the remaining 28%.

Reference Books:
Data Science from Scratch, Joel Grus, O’Reilly Media, 9781491901427
Advanced Analytics with Spark: Patterns for Learning from Data at Scale, Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills, O’Reilly Media; 1 edition, 9781491912768
Academic Honesty Statement:

Academic honesty plays a key role in our efforts to maintain a high standard of academic excellence and integrity. Students are advised that ALL acts of intellectual dishonesty will be handled in accordance with the SFU Academic Honesty and Student Conduct Policies (http://www.sfu.ca/policies/gazette/student.html).