ICS 438 / ICS 691B

Big Data Analytics

A. Overview

Data collection and generation have become easier and cheaper. For example, a person in a high-income country generates approximately megabytes of data every second. Despite technological advances on many fronts, “we are drowning in information, but starving for knowledge.”[1] Many factors might be contributing to this situation. Still, perhaps the most relevant is the lack of data-savvy professionals who can help derive scientific, economic, or social value from these data.

Data analytics presents various challenges, ranging from ingestion and management to analysis, visualization, and communication. This course will cover several big data concepts, with a special emphasis on text data. Its main objective is to introduce you to practical approaches and tools for processing, analyzing, and visualizing large datasets.

B. Course Objectives

1. Discover big data platform formats and specifications.
2. Learn fundamental algorithms and techniques to explore big data and generate actionable insights.
3. Getting familiar with machine learning and AI-based models for extracting insight from text data.

C. Course Objectives

The topics we will cover are:

1. Analyzing a problem to determine whether and how big data techniques can be applied.
2. Understand basic data mining techniques.
3. Learn how to adapt and extend computationally intensive data mining techniques to massive data sets.
4. Plotting and communicating big data and their insight.
5. Using deep learning models as a black box to extract features from text data
6. Implement a simple big data analytics solution in Python.

[1] Rutherford (Rudy) D. Rogers, Director of Major Research

D. Prerequisites

Students are expected to have obtained reasonable programming experiences. Although not required, knowledge of Python is a plus.

E. Instructor

Mahdi Belcaid
E-mail: mahdi@hawaii.edu
Office: 306B
Office hours: Hybrid Office hours through Zoom (Mondays 2:30-4:30)
TA: TBD
Course Website: TBD
Miro Board: TBD
Slack workspace: TBD

F. Teaching Assistant

Akib Sadmanee
sadmanee@hawaii.edu
Office hours: TBD

G. Lectures

Tuesday and Thursday
10:30- 11:45 PM
KUY213

H. Textbooks

References:

1. Tunstall, Lewis, Leandro von Werra, and Thomas Wolf. Natural language processing with transformers. " O'Reilly Media, Inc.", 2022.
2. Rajaraman, Anand, and Jeffrey David Ullman. Mining of massive datasets. Cambridge University Press, 2011.
3. VanderPlas, Jake. Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.", 2016. https://jakevdp.github.io/PythonDataScienceHandbook/

I. Grading and Student Evaluation

Grading will be as follows:

- A:≥90%
- B:≥80% and < 90%
- C:≥70% and < 80%
- D:≥60% and < 70%
- F:<60%

1 Assignments (45%)

Three assignments will be given. These will be turned in as Jupyter notebooks (http://jupyter.org/) containing the assignment code, as well as proper annotations, comments justifying the approach and, if necessary, graphs summarizing the results. Assignments using libraries not discussed in class will not be accepted. Points will be deducted when the specifications provided in the assignment are not met. The assignments will be submitted to the assignment GitHub project before 11:59 PM (HST) on the due day. Late work will be accepted with a 10% grade penalty for <24 hours of lateness and a 50% grade penalty for <48 hours of lateness. Assignments turned in after 48 hours will not be accepted.

2- Final Project – ICS438 only (25%)

ICS Students will be required to contribute a final project on a big data analytics topic of their choice. The final project proposed should use concepts from the three high-level topics covered in class, i.e., data processing, analysis, and visualization. Students will be graded on originality, relevance of questions, code quality, and appropriateness of proposed solution.

3. Final exam – ICS438 only (25%)

A comprehensive exam that assesses global understanding of the material and general concepts.

4. Attendance and Participation (5%)

A maximum of four scheduled classes may be missed, except when a valid justification is offered (e.g.: illness, religious holidays, and personal emergencies). Students must submit an online honor form with a unique code at the beginning of each session to confirm their attendance. Additionally, students are graded on their participation on the course Slack channel (e.g., answering questions, providing feedback, or sharing about course material and relevant topics).

5. Paper Review – ICS691 only (50%)

Students should complete an applied review of a big data research topic in the form of a tutorial. The review should examine methods related to a big data topic, preferably ones not covered in class, and should present the methods in a practical format in a preconfigured Docker environment with code samples and scripts and commands. The requirements will be provided in week four.

J. Academic Dishonesty

All occurrences of academic dishonesty, as defined below, will result in a grade of 0 for the assignment or exam, and in a memo in your ICS department file describing the incident. Which will be done for all students involved. Should there be more than one memo of this type in your file, the incident will be referred to the Dean of Students. Disciplinary sanctions range from a warning to expulsion from the university, as seen at:
http://www.catalog.hawaii.edu/about-uh/campus-policies1.htm

See relevant excerpts below:

Academic Integrity

The integrity of a university depends upon academic honesty, which consists of independent learning and research. Academic dishonesty includes cheating and plagiarism. The following are examples of violations of the Student Conduct Code that may result in suspension or expulsion from UH Manoa.

Cheating

Cheating includes, but is not limited to, giving unauthorized help during an examination, obtaining unauthorized information about an examination before it is administered, using inappropriate sources of information during an examination, altering the record of any grade, altering an answer after an examination has been submitted, falsifying any official UH Manoa record, and misrepresenting the facts in order to obtain exemptions from course requirements.

Plagiarism

Plagiarism includes, but is not limited to, submitting, to satisfy an academic requirement, any document that has been copied in whole or in part from another individual's work without identifying that individual; neglecting to identify as a quotation a documented idea that has not been assimilated into the student's language and style; paraphrasing a passage so closely that the reader is misled as to the source; submitting the same written or oral material in more than one course without obtaining authorization from the instructors involved; and "dry-labbing," which includes obtaining and using experimental data from other students without the express consent of the instructor, utilizing experimental data and laboratory write-ups from other sections of the course or from previous terms, and fabricating data to fit the expected results.

Disciplinary Action

The faculty member must notify the student of the alleged academic misconduct and discuss the incident in question. The faculty member may take academic action against the student as the faculty member deems appropriate. These actions may be appealed through the Academic Grievance Procedure, available in the Office of Judicial Affairs. In instances in which the faculty member believes that additional action (i.e., disciplinary sanctions and a UH Manoa record) should be established, the case should be forwarded to the Office of Judicial Affairs.