CS 4984 – Data Science and Analytics Capstone

Course Description

Researchers across disciplines are excited by the prospect of "data-driven science". This advanced project-based course is geared towards deriving valuable insights from data. Students are expected to integrate software engineering and data analytics skills acquired in previous courses. Team-based capstone data project will work on real-world challenges while drawing from areas of expertise within the department, e.g, text analytics, graph analytics, computational social science, data mining and visualization, machine learning, etc.

Course Overview

The principle aim of this capstone-style course is to develop insights from data. For Spring 2019, the focus will be on data generated via online social media platforms, such as, Facebook, Twitter, Reddit and the real-world challenges that emerge on these platforms (examples include, misinformation, hate speech, polarization, etc.). The first few weeks of this course will comprise multiple readings, in-class discussions, and in-class practicum sessions to introduce you to basic concepts of analyzing data left behind in social media platforms. During this time, you will have the opportunity to read technical papers, write your reflections where you will not just summarize the paper but think about what additional questions the paper enables. This is your chance to come up with a cool project idea based on what you just read. I will also provide you with a list of high level topics and suggestions. You will blog about your ideas, which will ultimately lead to team pitches and your project proposal. We will also have mid-term check points for your final projects and multiple practicum sessions during the course of the semester to get you warmed up for the main project by analyzing real-world social data.

Prerequisites

CS 3654 or equivalent, or permission of the instructor.
In terms of the required skills, students need to have basic knowledge of statistics and preliminary machine learning. An overview of the concepts and tools needed will be reviewed in class, however in-depth coverage of the fundamentals is not in the scope of this course. Students also need to be proficient in programming, especially using Python. Experience in use of a scientific computing software like R is a bonus, but not required. Students should be prepared to apply what they have learned in prior courses (like algorithms, computational thinking, etc.) to this emerging new field. You are expected to quickly learn many new things. For example, your project may require you to fetch Twitter data using the Twitter API or analyze posts from Reddit using pre-existing libraries (like python *nltk, sklearn*), which should not be too challenging if you already know high-level languages like Python. Please make sure you are comfortable with this.

Grading Criteria (Tentative)

- Class participation – 10%
- Reading responses – 10%
- Warm-up homework – 10%
- Data Practicums – 10%
- Term project – 60%
    - Project pitch presentation - 5%
    - Project proposal – 5%
    - Midterm project presentation - 5%
    - Midterm Report - 10%
    - Final project presentation - 10%
    - Final report - 25%

Topics (Tentative)

- Data collection
- Data pre-processing & cleaning
- Data exploration & visualization
- Basic text analysis
- Text classification
- Document clustering & classification
- Basic machine learning models and applications
- Hypothesis generation
- Experiments