ISE 467/ISE 367
Fall 2019

Mining of Massive (Large) Datasets

Dr. Martin Takáč
Mohler 481, Tuesday after lecture
takac@lehigh.edu

Suresh Bolusani
Mohler, office hours TBD
bsuresh@lehigh.edu

1. Course Information

Meeting Times:
Tuesday 9:20 am – 12:00
Thursday 10:45 am – 12:00
Location: Mohler Lab 121
Prerequisites:

2. Scope of the Course

Big Data is transforming the world! In today's digital world there is an ever increasing demand for solving "Big Data" problems, each described by gigabytes or terabytes of data from sources such as twitter feeds, online image databases, text corpora, videos, government records, scientific experiments or click behaviour of online users. Big Data mining is the capability of extracting useful information from these large datasets or streams of data. The Big Data challenge is becoming one of the most exciting opportunities for the next years.

In this course, we will explore how big data can be extracted and analyzed to discover new information that complements our existing knowledge of the system being studied. This will include a discussion of suitable algorithms for high dimensional data, graphs, machine learning. We introduce modern distributed programming model **MapRecue** and **Apache Spark** engine for large-scale data processing. We will learn how to work with large datasets on clusters and clouds (e.g. Amazon Clouds). You will learn how Google's PageRank algorithm models importance of Web pages and also many extensions that have been used. We will cover locality-sensitive hashing (how to find a similar pairs in such a large datasets that you cannot compare all pairs); Nearest Neighbors algorithm; we will discuss the importance of large graphs and how to mine them; when data is stored as a very large, sparse matrix, dimensionality reduction is often a good way to model the data, but standard approaches do not scale well – hence, we will talk about efficient approaches for dimension reduction. In the remaining part of course we will learn some aspects of collaborative filtering and clustering.

3. Class Methodology

In this class, we will be focusing on enhancing your knowledge, creativity, team work and entrepreneurial skills. The learning will be achieved mostly by working on projects (4-6 in total – observe that 40% of the grade is based on projects!).

In each project, you should identify the hidden information in the data, which can be used to improve current business (gain more profit). Each project will involve a short presentation of your findings (where you should communicate your findings).

4. Topics Covered in the Course

1. Introduction to Massive Datasets
2. Basics of Python/Scala
3. Map Reduce; Apache Spark Framework
4. Google Page Rank (Link Analysis)
5. Locality-Sensitive Hashing
6. Nearest Neighbors
7. Analysis of Large Graphs
8. Collaborative Filtering
9. Clustering
10. Review

5. Course Logistics

5.1 Textbook

`http://infolab.stanford.edu/~ullman/mmds/bookL.pdf`

5.2 Homework

There will be several homework, and **all must be completed to receive a grade for the course**.

Homework will be penalized for each day they are late. After solutions are released, they will not be accepted. No exceptions. Also, no exception to the no-exception rule.

5.3 Re-grade Requests

If you disagree with the grade you received on a homework or exam problem, you may submit a request for that problem to be re-examined. This request must be turned in no more than 48 hours after you receive the graded assignment. Once we re-examine your work and decide whether to change your grade, our decision will be final.

5.4 Class Preparation and Participation

You are expected to come to class regularly and to be prepared for each class by reading the relevant sections of the textbook ahead of time. I will post slides on Coursesite in advance so that you may bring them to class if you wish. In addition, you are expected to participate in class discussions and ask questions when you are confused. A portion of your grade will be based on class participation.

5.5 Extended Absences

If you believe you will miss two or more consecutive lectures due to illness, family emergencies, etc., please contact me as early as possible so that we can develop a plan for you to make up the missed material. Under no circumstances will I give credit for missed homework or exams unless you have discussed your absence with me sufficiently in advance.

5.6 Evaluation

Homework	15%
Midterm quiz	15%
Projects	40%
Final exam	20%
Class participation (discretion of the instructor)	10%

6. Plagiarism Policy

I strongly encourage you to consult with your colleagues when you're working on homework. However, you will not understand the material thoroughly or do well on the exams unless the work that you turn in is ultimately your own. Therefore, you must write up your answers alone, and without looking at anything you wrote down while working with your group. This means that if you solved the problem with a friend, you're going to have to go home and solve it all over again, by yourself. If you wrote code with a friend, you're going to have to re-write it by yourself. The work you turn in must be your own. In your write-up, you must cite everyone with whom you worked or consulted about each problem, as well as any books or other references that you used to solve the problem. Any breach of this policy will be considered an act of plagiarism, and no credit will be given for such assignments. Repeat offenses will be grounds for failure for the course.

7. Policies for the Course

- You are expected to arrive on time, turn your cell phone off, refrain from reading the newspaper, refrain from text-messaging the rest of the world, and stay in class for the duration of the lecture. Students who need to leave early should notify me ahead of time.
- Each homework in the course must be completed on its due date. Homework are due at the beginning of class.
- Any kind of cheating in any part of the course will be severely sanctioned and might result in disciplinary action.
- Regular attendance is required for the lectures. You should let me know in advance if you are going to be absent for a job interview, an athletic event, a religious holiday, a field trip, or any other good reason. Being sick is a good reason too, but you need to email me.
- If you plan to miss the lecture on a day where an assignment is due, you should make arrangements ahead of time so that your assignment is turned in on or before the due date. You are very strongly discouraged to miss a quiz.
- The lectures will be a lot more enjoyable if you participate.
- You are expected to check the course webpage regularly.
- Taping lectures, and specifically audio recording, is illegal in Pennsylvania without the prior consent of **all parties** in attendance.

8. Other

8.1 Accommodations for Students with Disabilities

If you have a disability for which you are or may be requesting accommodations, please contact both your instructor and the Office of Academic Support Services, University Center C212 (610-758-4152) as early as possible in the semester. You must have documentation from the Academic Support Services office before accommodations can be granted.

8.2 The Principles of Our Equitable Community

Lehigh University endorses The Principles of Our Equitable Community (http://www4.lehigh.edu/diversity/principles). We expect each member of this class to acknowledge and practice these Principles. Respect for each other and for differing viewpoints is a vital component of the learning environment inside and outside the classroom.