Course:

Time: Tuesdays and Thursdays 2:45PM – 4:05PM
Place: HU 113
Blackboard: The course uses the Blackboard Learning System. All materials and assignments will be handled there. Login at https://blackboard.albany.edu,

Instructor:

Instructor: Petko Bogdanov
Office**:** LI-95I, in the library building underground branch
Office Hours: Monday 1pm-2pm, Wednesday 10am-11am or by appointment
Email: **please use Blackboard mailing system**, usually you can rely on a 24-hour turnaround on your questions, as the account will be checked daily.

Course Topics

- MapReduce and Hadoop
- Frequent itemsets and Association rules
- Near Neighbor Search in High Dimensions
- Locality Sensitive Hashing (LSH)
- Dimensionality reduction: SVD and CUR
- Recommender systems
- Clustering
- Analysis of massive graphs
- Analysis: PageRank, HITS
- Web spam and TrustRank
- Proximity search on graphs
- Large-scale supervised machine learning
- Mining data streams
- Web advertising
- Optimizing submodular functions
- Anomaly Detection
- Distributed graph processing
- …More Advanced Topics

Reading assignments will be posted on Blackboard and announced in class. Reading for the class will be from:

- BOOK: Mining Massive Datasets by J. Leskovec, A. Rajaraman, J. Ullman (PDFs at http://mmds.org).
- Research papers for advanced topics

Goals: By the end of this course, you will:

- Be familiar with important DM problems
- Be able to use computing platforms such as MapReduce to mine large datasets
- Have basic knowledge of research principles in the domain of DM with emphasis on scale

Grading and Evaluation

40% - Final Project
40% - Homework
10% - In-class Quizzes
10% - Critical Paper Reviews, extra assignments
Extra Credit: up to 5% for in-class participation

Project

Teams: 1 or 2 members (No Exceptions)
Milestones: Unless otherwise stated, all milestones are due at midnight of the designated date.

- (Jan 28) Project groups + web site. Create a public website for the project containing the team members.
- (Feb 15) Project proposal (post on the project web site). This includes problem formulation, related work and how the project is different from the related work.
- (Feb 16) Flash presentation (due in class). 2 minutes per project in class presentation/advertisement of the proposed project (you can use a slide if you email it to the instructor the previous day)
- (Feb 15 - Feb 24) Come to office hours as a team to discuss your project.
- (Feb 29) Evaluation plan (update project website): A planned outline of datasets, what are you going to measure in order to evaluate the project.
- (Mar 20) Mid-project report (update project website): draft of your final write-up for the project. Even though some experiments/implementation might be missing. A special section on key risks/unknowns.
- (Apr 24) First experimental figure due
- (May 3) Final project presentation (due in class). Conference-style presentation with Q&A.
- (May) Project paper due in Blackboard. You are expected to use the ACM format to write your project reports (8 pages maximum, 4 pages minimum, including references; this page limit is strict).

Project report structure

1. Introduction/Motivation/Problem Definition: What is it that you are trying to solve/achieve and why does it matter?
2. Prior Work: How does your project relate to previous work? Please give a short summary on each paper you cite and include how it is relevant.
3. Model/Algorithm/Method: detailed description of your primary contribution. Clear and well including notation and any analysis of complexity, data structures, etc.
4. Results and findings: How do you evaluate your solution to whatever empirical, algorithmic or theoretical question you have addressed and what do these evaluation methods tell you about your solution? It is not so important how well your method performs but rather how interesting and clever your experiments and analysis are. Strive for a clear and conclusive experiments. Make sure to interpret the results and talk about what can be learned from them.

Note: Good writing, grammar, organization and figures are essential. Regarding writing, make sure you go over The Elements of Style. (http://www.bartleby.com/141/)

Policies

Late Turn-ins: homework turned in before or on the specified due date and time, in class or submitted through Blackboard, depending on the circumstance, are eligible for 100% of the grade. If you choose to turn in after the due date and time passes, for the first 24 hour period after the due date and time, your assignment will be eligible for 67% of the full grade; for the second 24 hour period after the due date and time, your assignment will be eligible for 33% of the full grade; for the third 24 hour period or later after the due date and time, your assignment will be eligible for 0% of the full grade.

Students with Disabilities: Students who feel that they have disabilities that require special arrangements for them to take the course must register with the Disability Resource Center (http://www.albany.edu/disability/index.shtml) Students are eligible for special services to which both the Center and the professor agree. In general, it is the student's responsibility to contact the professors at least one week before the relevant assignment to make arrangements.

Academic Integrity: The university's policies on academic integrity are listed here: http://www.albany.edu/undergraduate_bulletin/regulations.html. You will be held to these policies VERY STRICTLY WITHOUT EXCEPTIONS. You are expected to be familiar with them. Any incident of academic dishonesty in this course, no matter how "minor" will result in (a) No credit for the affected assignment, (b) a written report sent to the appropriate University authorities (e.g. the Dean of Undergraduate Studies) and a referral of the matter to the University Judicial System for disposition, (c) a final mark reduction by at least a full letter grade or in the severe case, an F grade of the course.

Assignment submission: All students should submit their homework solutions, as well as their final project reports in Blackboard. The submissions should involve two files:

1) Solution.pdf - Containing the homework solutions or project report, and
2) Code.zip - An archive of code for programing assignments and project code

The format for solutions is PDF. If you are using a word processor such as those in MS Word or Open Office, save the final solution as a PDF and submit only the pdf. Students can also scan their hand-written homework and submit them again as PDF. In Code.zip put all the code for a single question into a single file and upload it.

Prerequisites

Students are expected to have the following background:

- Good knowledge of Java since most assignments will involve the use of Hadoop/Java.
- Familiarity with basic probability theory – see refresher document in Blackboard
- Familiarity with writing rigorous proofs - see refresher document in Blackboard
- Familiarity with basic linear algebra - see refresher document in Blackboard
- Familiarity with algorithmic analysis

Important:

- Log in the Blackboard system and access the class space
- Meet your classmates and form teams