Course Information **Spring 2023**

CS246: Mining Massive Data Sets

Instructor: Jure Leskovec

Co-Instructor: Mina Ghashami

Lectures: 3:00PM - 4:20PM Tuesday and Thursday in Nvidia Auditorium, Huang Engineering Center

Course website: https://cs246.stanford.edu

Contacts:

- Use Ed to post questions: https://edstem.org/us/courses/38149/discussion/
- For external enquiries, personal matters, or in emergencies, e-mail us at cs246-spr2223-staff@lists.stanford.edu
- SCPD students can attend office hours remotely via zoom; we will use QueueStatus https://queuestatus.com/queues/2268 for maintaining the queue and the zoom links will be posted during the first week.

Course Coordinator: Lata Nair

Office Hours: refer to the course website for times and locations; announced by the end of week 1.

Topics

- MapReduce and Spark
- 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
- Link Analysis: PageRank, HITS
- Web spam and TrustRank
- Proximity search on graphs
- Large-scale supervised Machine Learning
- Mining data streams
- Learning through experimentation
- Web advertising
- Optimizing submodular functions

Assignments and grading

- 4 homework assignments requiring coding and theory (40%)
- Final exam (30%)
- Weekly Colab notebooks (30%)
- Extra credit: Ed and course participation, reporting bugs in course materials (up to 2%)

Homework policy

Questions We try very hard to make questions unambiguous, but some ambiguities may remain. Ask (i.e., post a question on Ed) if confused, or state your assumptions explicitly. Reasonable assumptions will be accepted in case of ambiguous questions.

Honor code We take honor code extremely seriously: (https://communitystandards.stanford.edu/policies-and-guidance/honor-code). The standard penalty includes a one-quarter suspension from the University and 40 hours of community service. We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down the solutions and the code independently. In addition, each student should write down the set of people whom s/he interacted.

Late assignments Each student will have a total of 2 late periods to use for homework assignments. Homework are due on Thursdays and late periods extend to 11:59 PM on the following Monday. No assignment will be accepted more than one late period after its due date.

Assignment submission All students (SCPD and non-SCPD) submit their assignments via Gradescope (http://www.gradescope.com). Students can typeset or scan their homework.
Students also need to upload their code on Gradescope. Put all the code for a single question into a single file and upload it. Refer to the course FAQ for more info.

Regrade requests We take great care to ensure that grading is fair and consistent. Since we will always use the same grading procedure, any grades you receive are unlikely to change significantly. However, if you feel that your work deserves a regrade, submit your request within a week of receiving your grade on Gradescope. However, note that we reserve the right to regrade the entire assignment.

Colabs Colab notebooks are posted on Thursdays and due one week later (hard deadline Thursday 11:59pm Pacific time). Once the deadline has passed, students will not be able to submit the results of their Colabs on Gradescope.

Prerequisites

Students are expected to have the following background (recitation sessions will refresh these topics):

- The ability to write non-trivial computer programs (at a minimum, at the level of CS107). Good knowledge of Python/Java will be extremely helpful since most assignments will require the use of Spark.
- Familiarity with basic probability theory is essential (at a minimum, at the level of CS109 or Stat116).
- Familiarity with writing rigorous proofs (at a minimum at the level of CS 103).
- Familiarity with basic linear algebra (e.g., any of Math 51, Math 103, Math 113, CS 205, or EE 263).
- Familiarity with algorithmic analysis (e.g., CS 161).

Materials

Notes and reading assignments will be posted on the course Web site. Readings for the class will be from:

- Mining Massive Datasets by J. Leskovec, A. Rajaraman, J. Ullman (PDFs at http://mmds.org).

Important dates

Assignment	Out Date	Due Date (11:59 PM)
Homework 1	Apr 6	Apr 20
Homework 2	Apr 20	May 4
Homework 3	May 4	May 18
Homework 4	May 18	Jun 1
Final exam		Jun 6 (Tentative)

We will also hold three review sessions in the first two weeks (sessions will be video recorded; dates TBD):

- Spark tutorial.
- Probability and proof techniques.
- Linear algebra.