Course Information **Winter 2019**

CS246: Mining Massive Data Sets

Instructor: Jure Leskovec
Office Hours: Tuesdays 9-10AM, Gates 418

Co-Instructor Michele Catasta
Office Hours: Thursdays 5-7PM, Gates 452

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

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

Contact:

- E-mail us at cs246-win1819-staff@lists.stanford.edu
- Use Piazza to post questions: http://piazza.com/stanford/winter2019/cs246
- SCPD students can attend office hours remotely via videoconferencing; the link will be posted on Piazza just before the office hours start.

TAs and office hours: See the course website for times and locations.

Topics

- MapReduce and Spark/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
- 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

- Homework 0 and Four problem sets requiring coding and theory (40%)
- Final exam (40%)
- Weekly Gradiance quizzes (20%)
- Extra credit: Piazza and course participation, reporting bugs in course materials (up to 1%)

Homework policy

Questions We try very hard to make questions unambiguous, but some ambiguities may remain. Ask (i.e., post a question on Piazza) 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 homeworks. Homework are due on Thursdays and late periods extend to midnight (11:59PM) 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 homeworks via Gradescope (http://www.gradescope.com). Students can typeset or scan their homeworks.
Students also need to upload their code at http://snap.stanford.edu/submit. 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.

Gradiance Quizzes are posted on Tuesday afternoon and due 9 days later (hard deadline Thursday 11:59pm Pacific time). Once the deadline has passed students will not be able to submit the quiz.

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 Hadoop/Java.
- 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 (all 23:59pm)
Spark tutorial	now	Jan 24
Assignment 1	Jan 10	Jan 24
Assignment 2	Jan 24	Feb 7
Assignment 3	Feb 7	Feb 21
Assignment 4	Feb 21	Mar 7
Final exam		Mar 19, 3:30-6:30pm

We will also hold three review sessions in the first two weeks of the course (sessions will be video recorded):

- Spark tutorial and help session. Thursday, January 10, 4:30-5:50 PM, Location TBD.
- Review of basic probability and proof techniques. Tuesday, January 15 4:30-5:50 PM, Location TBD.
- Review of basic linear algebra. Thursday, January 17 from 4:30-5:50 PM, location TBD.

Next steps for students

- Register for Piazza: http://piazza.com/stanford/winter2019/cs246
- Register for Gradiance: http://www.newgradiance.com/services class token 3DBCAD12
- Register for Gradescope: https://gradescope.com/ course code MNPBKE
- Download Spark VM, start the tutorial: http://cs246.stanford.edu/homeworks/hw0.tar.gz