Course Information **Spring 2020**

CSE 547: Machine Learning for Big Data

Instructor: Tim Althoff

Lectures: Tuesday/Thursday 10:00-11:20am PDT via Zoom. You can find links to the Zoom lectures on Canvas (https://canvas.uw.edu/courses/1372006/external_tools/95443).

Course website: https://courses.cs.washington.edu/courses/cse547/20sp/

Contact:

- Use EdDiscussion to post questions: https://us.edstem.org/courses/422/discussion/.
- For emergencies and personal questions, e-mail us at cse547-instructors@cs.washington.edu

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

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
- Optimizing submodular functions

Assignments and grading

- **Homeworks** (40%): Four problem sets requiring coding and theory (10% each)
- **Final project** (40%)
- **Colabs** (20%): 10 colabs in total, released weekly (2% each)
- **Extra credit**: EdDiscussion 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 EdDiscussion) 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://www.cs.washington.edu/academics/misconduct). We strongly encourage students to form study groups. Students may verbally discuss and work on homework problems in groups. However, each student must write down the solutions and the code independently. Students may not share written work or programs (on paper, electronic, or any other form) with anyone else. Importantly, in submissions, each student should write down the set of people whom they interacted with including anyone not taking this class or not working at UW (excluding the instructor and TA(s)). Students should appropriately cite any helpful material they find in published literature or on the web including assignment's answer, or partial answer. Students should not claim to have come up with an idea that wasn't originally theirs; instead, should explain it in their own words and make it clear where it came from.

Late assignments Each student will have a total of two late periods to use for homeworks. A late period lasts 48 hours from the original homework deadline. No assignment will be accepted more than one late period after its due date.

Assignment submission All students must submit their homeworks via Gradescope (http://www.gradescope.com). Students can typeset or scan their homeworks. Scanned homeworks must be absolutely clear and legible. Points may be deducted for unclear parts of handwritten 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.

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 CSE 332, CSE 373, or equivalent). 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 (any introductory probability course).
- Familiarity with writing rigorous proofs (e.g. CSE 311 or equivalent).
- Familiarity with basic linear algebra (e.g. MATH 308).
- Familiarity with algorithmic analysis (e.g. CSE 417, CSE 421).

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)
Colab 0, Colab 1	Apr 2	Apr 9
Assignment 1	Apr 2	Apr 16
Colab 2	Apr 9	Apr 16
Project proposal		Apr 23
Colab 3	Apr 16	Apr 23
Assignment 2	Apr 16	Apr 30
Colab 4	Apr 23	Apr 30
Project milestone		May 7
Colab 5	Apr 30	May 7
Assignment 3	Apr 30	May 14
Colab 6	May 7	May 14
Colab 7	May 14	May 21
Assignment 4	May 14	May 28
Colab 8	May 21	May 28
Colab 9	May 28	Jun 7
Final report		Jun 7
Final presentation		Jun 8

We will also hold three review sessions in the first two weeks of the course:

- Spark tutorial and help session. Thursday, Apr 2, 1:00-3:00pm.
- Review of basic probability and proof techniques. Tuesday, Apr 7, 3:30-5:30pm.
- Review of basic linear algebra. Thursday, Apr 9, 1:00-3:00pm.

Next steps for students

- Register for EdDiscussion: https://us.edstem.org/courses/422/discussion/
- Register for Gradescope: https://www.gradescope.com/courses/95372 with entry code: MP8KGN
- Register for Canvas: https://canvas.uw.edu/courses/1372006/
- Start planning for the course project. Sign up with your team here: https://forms.gle/oG8ckShER5yDHHs37
- Complete Colab 0/1 released on Thursday