

CS426 COURSE OVERVIEW

Spring 2017



GENERAL INFORMATION

- Instructor: Yao SHEN (沈耀)
 - Email: [yshen AT cs.sjtu.edu.cn](mailto:yshen@cs.sjtu.edu.cn)
 - Office: SEIEE Building 3-535
- Course web site:
 - <http://www.cs.sjtu.edu.cn/~yshen/courses/BigData/>
- Teaching Assistant:
 - 沈国栋 gdshen@sjtu.edu.cn
 - 陈健 CJ528122401@hotmail.com



TEXTBOOK

- Anand Rajaraman and Jeffrey D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2011.

You can download it from the [book website](http://www.mmds.org/) (http://www.mmds.org/).



REFERENCE BOOKS

- Jiawei Han, and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Second Edition, 2006.
- Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- Chuck Lam. *Hadoop in Action*. Manning Publications, First Edition, 2010.
- Holden Karau, Andy Konwinshi, Patrick Wendell, Matei Zaharia, Learning Spark, O'REILLY, 2015.
- Nick Pentreath. *Machine Learning with Spark*. Packt Publishing, 2015.



COURSE TOPICS

-SUBJECT TO CHANGE

- Introduction : Data-Intensive Scalable Computing (DISC) & Data Mining
- Parallel & Distributed Computing (esp. Cloud Computing)
 - OpenMP, Pthreads, MPI
 - MapReduce (Hadoop) and Spark
- Data Mining and Machine Learning
 - Association rules, Latent semantic indexing, Dimensionality Reduction
 - Clustering, Supervised learning
- Data-Intensive Applications
 - Search, link analysis, recommender systems, advertising on Web
- Extra-Topics
 - From data processing companies



PREREQUISITES

- Data structure
- Design and analysis of algorithms
- Linear algebra
- Probability theory
- Programming languages : Java, c++



GRADING SCHEME

–SUBJECT TO CHANGE

- Homework (40%)
- Final exam (60%)



LATE ASSIGNMENTS

- Homework is due on the assigned date.
- Late submissions of homework or project will receive partial or no credit.
 - 20% penalty for per day.
 - NOT accepted 72 hours after the due date.



ACADEMIC HONOR CODE

- Honesty and integrity are central to the academic work.
- All your submitted assignments must be entirely your own (or your own group's).
- Any student found cheating or performing plagiarism will fail this course.

