

Data-Intensive Text Processing with MapReduce

Jimmy Lin and Chris Dyer

University of Maryland, College Park
{jimmylin,redpony}@umd.edu

1. Overview

This half-day tutorial introduces participants to data-intensive text processing with the MapReduce programming model [1], using the open-source Hadoop implementation. The focus will be on scalability and the tradeoffs associated with distributed processing of large datasets. Content will include general discussions about algorithm design, presentation of illustrative algorithms, case studies in HLT applications, as well as practical advice in writing Hadoop programs and running Hadoop clusters.

2. Intended Audience

The tutorial is targeted at any NLP researcher who is interested in data-intensive processing and scalability issues in general. No background in parallel or distributed computing is necessary, but a prior knowledge of HLT is assumed.

3. Course Objectives

- * Acquire understanding of the MapReduce programming model and how it relates to alternative approaches to concurrent programming.
- * Acquire understanding of how data-intensive HLT problems (e.g., text retrieval, iterative optimization problems, and graph algorithms) can be solved using MapReduce.
- * Acquire understanding of the tradeoffs involved in designing MapReduce algorithms and awareness of associated engineering issues.

4. Tutorial Topics

The following represents a tentative list of topics that will be covered:

- * Introduction to parallel and distributed processing
- * Introduction to MapReduce
- * Tradeoffs and issues in algorithm design
- * Simple counting applications (e.g., relative frequency estimation)
- * Applications to inverted indexing and text retrieval
- * Applications to graph algorithms

- * Applications to iterative optimization algorithms (e.g., EM)
- * Case study in machine translation
- * Tips and tricks in writing Hadoop programs
- * Practical issues in running Hadoop clusters

5. Instructor Bios

Jimmy Lin is an Associate Professor in the iSchool at the University of Maryland, College Park. He joined the faculty in 2004 after completing his Ph.D. in Electrical Engineering and Computer Science at MIT. Dr Lin's research interests lie at the intersection of natural language processing and information retrieval. He leads the University of Maryland's effort in the Google/IBM Academic Cloud Computing Initiative. Dr. Lin has taught two semester-long Hadoop courses and has given numerous talks and tutorials about MapReduce to a wide audience.

Chris Dyer is a Ph.D. student at the University of Maryland, College Park, in the Department of Linguistics. His current research interests include statistical machine translation, machine learning, and the relationship between artificial language processing systems and the human linguistic processing system. He has served on program committees for AMTA, ACL, COLING, EACL, EMNLP, NAACL, ISWLT, and the ACL Workshops on Machine translation, and is one of the developers of the Moses open source machine translation toolkit. He has practical experience solving NLP problems with both the Hadoop MapReduce framework and Google's MapReduce implementation, which was made possible by an internship with Google Research in 2008.

References

[1] Dean, Jeffrey and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI 2004), p. 137-150, 2004, San Francisco, California.

[2] Jimmy Lin. Exploring Large-Data Issues in the Curriculum: A Case Study with MapReduce. Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics (TeachCL-08) at ACL 2008, p. 54-61, 2008, Columbus, Ohio.