Data-Intensive Text Processing with MapReduce

Jimmy Lin and Chris Dyer
University of Maryland, College Park
{jimmylin,redpony}@umd.edu

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.

Amazon has generously agreed to provide each participant with $100 in Amazon Web Services (AWS) credits that can used toward its Elastic Compute Cloud (EC2) "utility computing" service (sufficient for 1000 instance-hours). EC2 allows anyone to rapidly provision Hadoop clusters "on the fly" without upfront hardware investments, and provides a low-cost vehicle for exploring Hadoop.

Intended Audience

The tutorial is targeted at any NLP researcher 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.

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, etc.) can be solved using MapReduce.
- Acquire understanding of the tradeoffs involved in designing MapReduce algorithms and awareness of associated engineering issues.

Tutorial Topics

The following lists topics that will be covered:

- MapReduce algorithm design
- Distributed counting applications (e.g., relative frequency estimation)
- Applications to text retrieval
- Applications to graph algorithms
- Applications to iterative optimization algorithms (e.g., EM)
- Practical Hadoop issues
- Limitations of MapReduce

Instructor Bios

Jimmy Lin is an assistant 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.

Proceedings of NAACL HLT 2009: Tutorials, pages 1–2,
Boulder, Colorado, June 2009.

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 [2] and has given numerous talks 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.

Acknowledgments

This work is supported by NSF under awards IIS-0705832 and IIS-0836560; the Intramural Research Program of the NIH, National Library of Medicine; DARPA/IPTO Contract No. HR0011-06-2-0001 under the GALE program. Any opinions, findings, conclusions, or recommendations expressed here are the instructors' and do not necessarily reflect those of the sponsors. We are grateful to Amazon for its support of tutorial participants.

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.