A Scalable Approach for Topic Modeling with R

Tiffany A. Connors
Texas State University
tiffanyaconnors@gmail.com

Ritu Arora
Texas Advanced Computing Center
rauta@tacc.utexas.edu

Motivation

R is available on several High Performance Computing (HPC) platforms that are accessible through the national cyber infrastructure (e.g., the Stampede and Wrangler supercomputers at the Texas Advanced Computing Center). By using R scripts in high-throughput computing mode, users can take advantage of the state-of-the-art HPC platforms and large-scale storage resources without any code rewriting.

Background

What is Topic Modeling?

Topic modeling is a form of text mining that uses probabilistic models to identify patterns of word use within a corpus. These topics can then be used to annotate the documents and to organize, summarize, and search document collections.

What is LDA?

Latent Dirichlet allocation (LDA) is a hierarchical Bayesian model that discovers semantic topics within text corpora. Topics are discovered by identifying groups of words in that frequently occur together within the documents. Each document in the collection can be assigned a probability of it belonging to a topic [1].

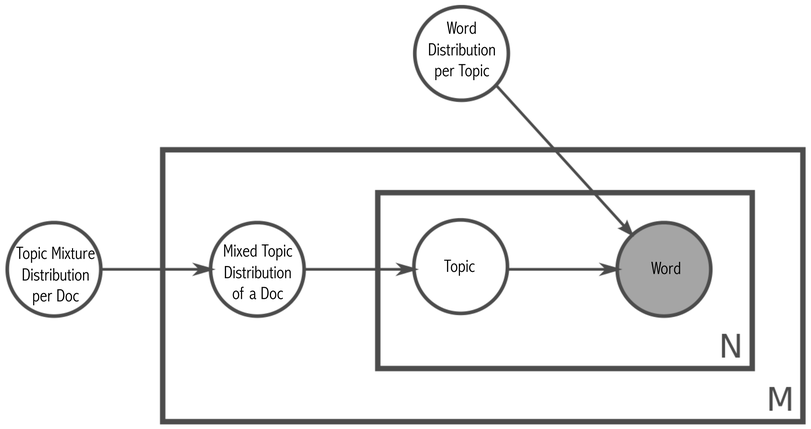


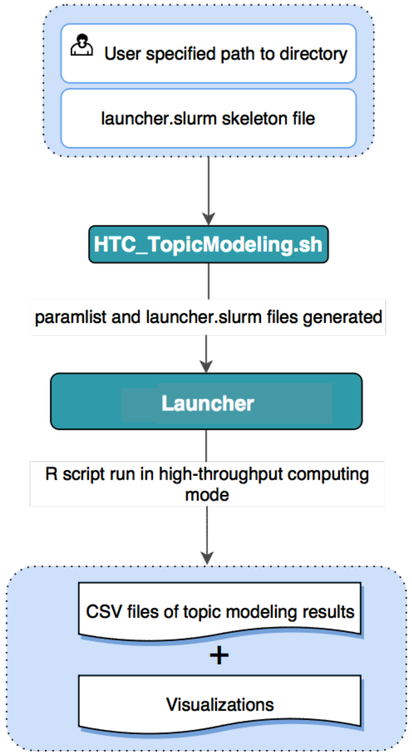
Fig 1. Graphical Representation of LDA. M, the outer box, is the documents in a corpus and N, the inner box, is the topics and words within a single document.

What is High-Throughput Computing (HTC)?

Multiple copies of a serial application can be run concurrently on multiple cores and nodes of a platform such that each copy of the application uses different input data or parameters. This leads to a reduction of the overall runtime and is called HTC. HTC mode can be used on the Stampede supercomputer at TACC by utilizing the Launcher tool that was developed in-house at TACC [5]. The Launcher requires the user to supply a SLURM job script (named "launcher.slurm") and a file containing the list of commands to run concurrently (named "paramlist").

Experimental Design

To demonstrate the effectiveness of utilizing HTC for R scripts, we used LDA topic modeling as a case study. We performed topic modeling on three publicly available datasets: BBC, BBCSport [5], and NSF Research Award Abstracts [6] for 1990 using the Stampede supercomputer at TACC.



Workflow Overview

1. HTC_TopicModeling.sh determines the number of subdirectories within the user specified directory.
2. The command for analyzing each subdirectory is added to the paramlist file.
3. The script checks for files to process in the top directory, if files exist, the directory is added to the paramlist
4. The number of cores is determined by the number of tasks in the paramlist.
5. If the number of tasks is greater than 16 (max number of cores/node), additional nodes are requested.
6. A custom launcher.slurm file is generated based on the needs of the job.
7. The job is submitted to the queue using sbatch.
8. The job is completed in parallel using HTC mode.

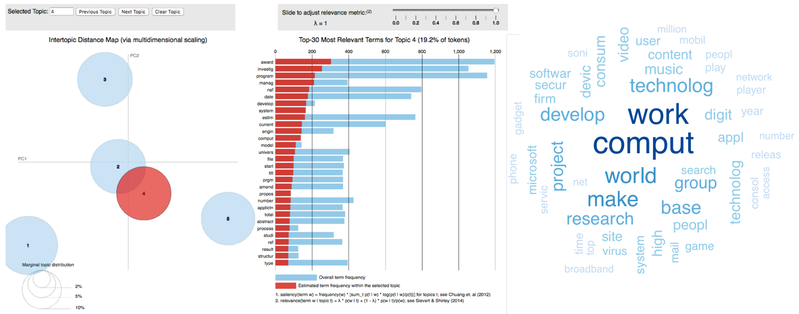


Fig 2. Sample Output. The interactive visualizations (left, center) and word cloud (right) generated by the R script. View the interactive visualization by scanning the QR code.

Acknowledgments

We are grateful for the support received from the National Science Foundation (NSF) funded ICERT REU program and to TACC and XSEDE for providing access to Stampede - a resource funded by NSF through award ACI-1134872. XSEDE is funded by NSF through award ACI-1053575.

Results and Discussion

By utilizing HTC, we were able to improve the runtime performance of the topic modeling script by nearly a factor of 3 for both BBC and BBCSports datasets.
In the case of the NSFAwards:1990 dataset, HTC improved the runtime by a factor of 23.

Dataset	Docs	Serial Time	HTC Time	Cores
NSFAwards:1990	10,097	114m 57s	4m 50s	42
BBC	2,225	21m 25s	7m 23s	5
BBCSport	737	10m 12s	3m 18s	5

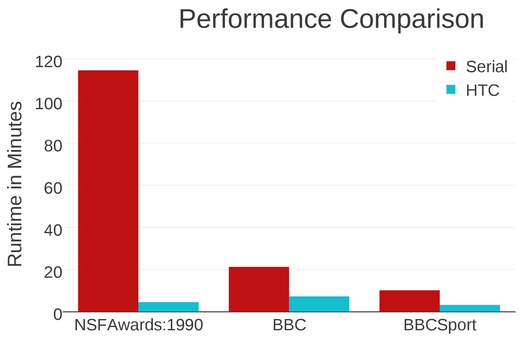


Fig 3. Comparison of Serial and HTC Runtimes. For all three datasets, the HTC processing performed notably better than the serial run.

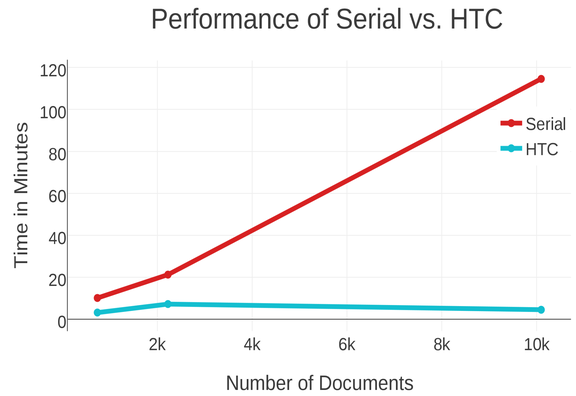


Fig 4. Performance of Serial vs. HTC. The execution times of processing each dataset are plotted in relation to the number of documents.

For classifying 10,097 documents, it took nearly 115 minutes while using a single core on the Stampede supercomputer. By porting the R scripts to Stampede and running them in the HTC mode using our HTC_TopicModeling script, we can classify these documents in under 5 minutes. The approach is scalable and can be generalized for course-grained data parallelism of other types of R scripts.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (March 2003), 993-1022.

[2] https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf

[3] C. Sievert and K. Shirley. Ldavis: A method for visualizing and interpreting topics. In 2014 ACL Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, June 2014.

[4] https://www.tacc.utexas.edu/research-development/tacc-software/the-launcher

[5] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.

[6] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.