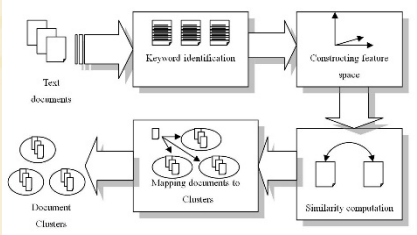
IMPLEMENTATION OF DISTRIBUTED MINIBATCH K-MEANS FOR TEXT CATEGORIZATION

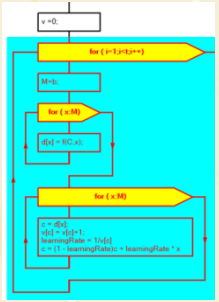
Rudrani Angira and Siddharth Jain

ABSTRACT

Clustering is widely used in machine learning domain. Lloyd's classic algorithm is not suitable for time constrained operations involving large data, because it is computationally expensive. An alternative to it is mini-batch k-means algorithm, which is significantly less computationally expensive and hence delivers quality results quickly.
We take the serial single machine mini-batch k-means algorithm and transform it into a faster distributed multi-threaded algorithm using Hadoop with Harp framework. We use RCV1-RCV2 dataset, consisting of 402207 documents for benchmarking purposes.



Algorithm : MiniBatch K-means



Parallel MiniBatch K-means for *n* iterations

DATA FORMAT

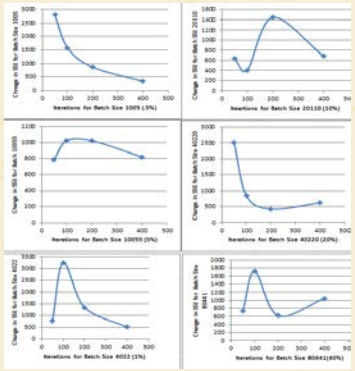
Each vector in a file of vectors is represented by a single line of the form:
<did> [<tid>:<weight>]+
where we have:
<did> : Reuters-assigned document id.

<tid> : A positive integer term id. Term ids are between 1 and 47,236.
<weight> : The numeric feature value, i.e. within document weight, assigned to this term for this document, as described in LYRL2004.

Example of the vector file format:
999995 1:0.03 3:0.047 8:0.3874973847893747 9
14:0.1 2748:0.03
999996 7:0.13 19:0.138 255:0.58588
314:0.28101 18800:0.005

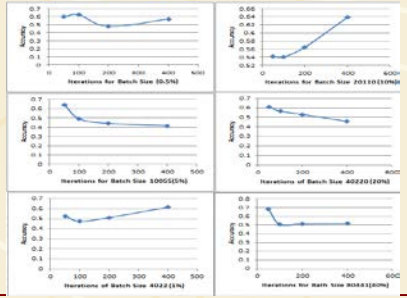
EXPERIMENT AND CLUSTER VALIDATION

SSE is used for determining the goodness of cluster, without any external information.



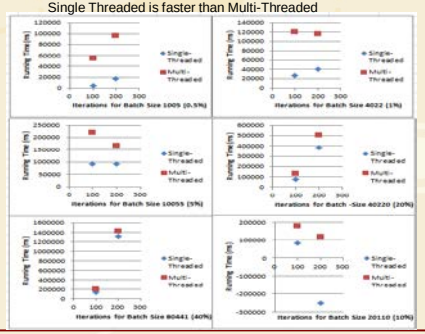
CATEGORIZATION ACCURACY

1. Documents most similar to the final centroids are found.
2. For every document in every cluster, we find the count of documents classified with at least one common base category as centroid representative document of the centroid it belongs to. This is an approximation and might not be the perfect way to measure the accuracy.



MULTI-THREADED vs SINGLE-THREADED

Single Threaded is faster than Multi-Threaded



ANOMALIES

Accuracy should improve with more no. of iterations, but it is not happening so in our experiments.
There are two possible reasons, because of which this anomaly is observed:
a. The choice of initial centroid can impact k-means.
b. Established centroids may be representing some other feature than the categories.

CONCLUSIONS

1.Mini-batch k-means can indeed be implemented in distributed fashion and can give good clusters (with respect to SSE).

2. These mini-batches mostly reduce the amount of computation required to converge to a local solution.

REFERENCE

[1] Yadav, M. K., & Baria, M. J. Mini-Batch Kmeans Clustering Using Map-Reduce in Hadoop. International Journal of Computer Science and Information Technology Research, 2(2), 336-342.

[2] Wan, J., Yu, W., & Xu, X. (2009, December). Design and implement of distributed document clustering based on MapReduce. In Proceedings of the Second Symposium International Computer Science and Computational Technology (ISCSCT), Huangshan, PR China (pp. 278-280).

[3] Lewis, D. D. RCV1-v2/LYRL2004: The LYRL2004 Distribution of the RCV1-v2 Text Categorization Test Collection (5-Mar-2015 Version). http://www.jmlr.org/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm

[4] Sculley, D., 2010 Web-scale k-means clustering, in: Proceedings of the 19th International Conference on World Wide Web. ACM, pp. 1177–1178.

INDIANA UNIVERSITY BLOOMINGTON
SCHOOL OF INFORMATICS AND COMPUTING