Acceleration of text data clustering using Xeon Phi processor

Marcin Pietroń, Jarosław Szymla, Maciej Wielgosz, Kazimierz Wiatr

Introduction

- Machine learning algorithms are widely used in many fields of scientific disciplines
- Most common usage of them is prediction, classification or data analysis
- Many hardware and software vendors like Intel, Cadence, NVIDIA etc. optimize machine and deep learning models by designing new hardware architecture or developing new libraries that can use efficiently resources available on specific hardware.
- In this work DAAL library was explored working on Xeon Phi Knights Landing hardware accelerator.

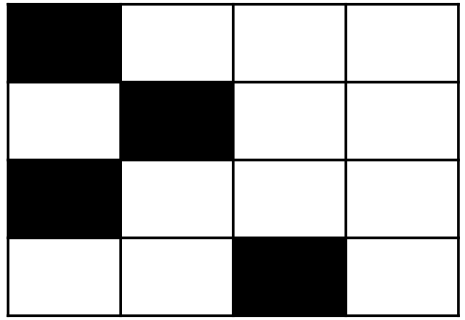
Description of a tests

- In our work we have used K-means clustering on a huge Wikipedia text dataset as one of the step of creating new binary model of text representation.
- The clustering was a main bottleneck in our algorithm. The algorithm produces retina which is binary representation of each word from corpus vocabulary.
- The first step of algorithm is to create VSM (vector space model) of each document in analyzed dataset.
- Then k-means algorithm is executed. Number of generated clusters is equal to resolution of retina. For each term from vocabulary number of documents in each cluster is computed in which given term occurs.

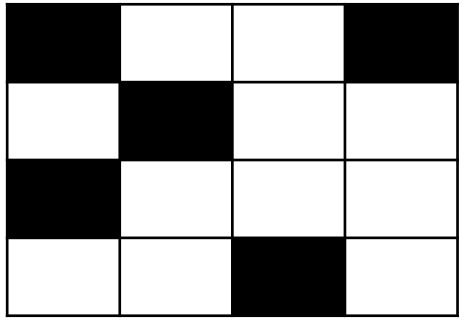
Description of a tests

- After that each pixel (assigned to one unique cluster) in term retina represents number of documents in which given token appears. The final stage is to set retina sparsity. This is done by setting threshold value and setting pixels with higher values to 1 and smaller to 0.
- The model can be efficiently processed in hardware accelerators like FPGA or DSP processors. Whole algorithm is implemented in python language using scikit learn and DAAL library.

Retina



cat



dog

Results

	Iterations	time	clusters	dataset size
CPU time	5	17518.3	1024	5000x4360000
Wall clock time (72 cores with 4 threads)	5	279.975	1024	5000x4360000

Results

Library	time per iteration	clusters	dataset size
scikit-learn (1 core)	113873,9	1024	5000x4360000
DAAL (72 cores)	39,41	1024	5000x4360000

Future work

- future work will concentrate on measuring efficiency of deep learning and genetic algorithms on Xeon Phi processor and compare it with modern general purpose graphic cards.
- further optimization of algorithm for generating semantic binary text model is considered.