# Installation

## Prerequisites

Java 8

Ant

Hadoop 2.6 with Harp

If you don’t have these follow the following guide to install the above:



## Code

There are two source code folders:

“miniBatchMultiThreaded” -> Source code for multi-threaded version

“miniBatchSingleThreaded” -> Source code for single-threaded version

Enter the desired folder and run ant in it.

Running ant will build the project and create build folder, consisting of “harp3-app-hadoop-2.6.0.jar”

Copy the newly created “harp3-app-hadoop-2.6.0.jar” in $HADOOP\_HOME.

# Data

Since the size of data exceeds 100 MB limit, it could not be uploaded on github.

The data is available on following IU Box link:

https://iu.box.com/s/lv07v8fefosuij2h9mnsouhp89mq9xoe

Download the whole folder “data”. Note the absolute path of the downloaded folder on your local system, this will be required at later stage for execution of application.

# Execution

The command line arguments for this application are:

<batchSizeInPercent> -> batch size in terms of percentage of total documents available. The input should be of type double. If required batch size is 10%, enter 10

<num of Centroids> -> num of centroids to form. The input type should be of type int.

<number of iteration> -> no. of iterations in mini-batch k-means. The input type should be of type int.

<workDir> -> Absolute path of Hadoop working dir

<localDir for data> -> Absolute path of local directory containing the data. Same directory which was downloaded in last step.

<localDir for application output> -> Absolute path of local directory where the output of application will be written

## Command format:

hadoop jar harp3-app-hadoop-2.6.0.jar edu.iu.km.KmeansMapCollective <batchSizeInPercent> <num of Centroids> <number of iteration> <workDir> <localDir for data> <localDir for application output>

## Command Example:

hadoop jar harp3-app-hadoop-2.6.0.jar edu.iu.km.KmeansMapCollective 10 4 200 /temp/term /N/u/user/data /N/u/user/output

# Output

An output file will be created on successful execution. The file will have a name in the form output-<time-stamp>

An output file will also be created on hdfs on the working path specified during invocation. The output file will be in the folder “appOutput1” with file name “result”